Visual Human-Computer Interaction

Stets, Jonathan Dyssel

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Visual Human-Computer Interaction

Jonathan Dyssel Stets

Kongens Lyngby 2018

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

PhD-2018-470
ISSN: 0909-3192

Summary (English)

Technologies such as Virtual and Augmented Reality has gained extensive popularity in the recent years. Simultaneously, vision systems and computing power have reached a point, where it is possible to acquire and process geometric and appearance data to produce photorealistic renderings that can appear indistinguishable from real photographs. This enables new ways for Human-Computer Interaction (HCI) methods and applications, which should be further evaluated to explore their full potential.

This thesis addresses a set of vision based challenges concerning HCI. The presented contributions fall into the overall themes of geometric acquisition of refractive objects, photorealistic rendering for computer graphics applications, and demonstration of systems for advanced and realistic applications for HCI. Accordingly, the work of this thesis is presented in a four-element taxonomy: Geometry and appearance digitization, tracking, visualization and interaction, and datasets. The work contributes to state of the art methods and prepares the ground for future research within the above-mentioned topics and generally the field of visual HCI.

Summary (Danish)

Teknologier som Virtual Reality og Augmented Reality har inden for de seneste år opnået massiv popularitet. Samtidig er visuelle computersystemer og beregningskraft nået til et punkt, hvor det er muligt at opsamle og processere geometri og komplekse materialemodeller til produktion af fotorealistiske renderinger, som for mennesker kan være svære at skelne fra fotografier. Dette skaber nye muligheder for metoder og applikationer for interaktion imellem mennesket og computeren (Human Computer Interaction - HCI), der derfor bør undersøges nærmere for at udforske deres fulde potentiale.

Denne afhandling adresserer Computer Vision-baserede problemstillinger relateret til HCI, og bidragene falder inden for følgende overordnede temaer: Digitalisering af geometrisk data og håndtering af refraktive (gennemsigtige) objekter, fotorealistisk rendering til applikationer inden for computergrafik samt systemer til avancerede og realistiske applikationer inden for HCI. Projekterne præsenteret i denne afhandling er opdelt i fire kategorier: Digitalisering af geometri og komplekse materialemodeller, kamerabaseret tracking, visualisering og interaktion samt datasæt. Bidragene er nye metoder inden for ovenstående emner - og overordnet set forbedringer til visuel baseret computerinteraktion (Visual Human Computer Interaction).

Preface

This thesis was prepared at the Image Analysis and Computer Graphics section at the Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU). It is done in fulfilment of the requirements for obtaining a doctor of philosophy degree in applied mathematics and computer science with emphasis on the field of computer vision.

The research presented in this thesis contributes with 7 publications within the topics of geometry and appearance digitization, tracking, and visualization and interaction.

The project has been supervised by Associate Professor Henrik Aanæs. The research has been carried out at the section for Image Analysis and Computer Graphics, DTU Compute. External research has been conducted at Massachusetts Institute of Technology and at the University of California, San Diego. Part of the research in this thesis has been in collaboration with Copenhagen Business School and the company iMotions. The PhD is partly funded by the Innovation Fund Denmark.

Lyngby, 5-March-2018

Jonathan Dyssel Stets

Acknowledgements

First of all, I would like to thank my supervisor Henrik Aanæs for valuable guidance throughout the project. I would also like to thank Rasmus Larsen and Ulrik Jensen for establishing this PhD project.

Thanks to collaborators at iMotions: Steen Christensen and Kåre L. Jensen, and to collaborators at Copenhagen Business School: Seidi Suurmets and Jesper Clement.

Thanks to Scott W. Greenwald and Pattie Maes for allowing me to visit the Fluid Interfaces Group at the MIT Media Lab. I learned a lot and met many great people during my stay here.

Thanks to Manmohan Chandraker for allowing me to visit Center for Visual Computing at University of California, San Diego. It was also a pleasure to meet and work with the people from this group, and I have learned a lot during this stay.

Thanks to friends and colleagues at the Image Analysis and Computer Graphics Section at DTU for both academic and non-academic discussions. A great thanks to my office mates Jannik Boll Nielsen, Eythor Runar Eiriksson and Jakob Wilm. And also thanks to project collaborators Jeppe Revall Frisvad, Alessandro Dal Corso, Rasmus Ramsbøl Jensen, Rasmus Ahrenkiel Lyngby, Sebastian Nesgaard Jensen, Andrea Luongo and Mads Doest.

Finally, a great thanks for the support from my family, friends and Helle Bumbech Andersen.

List of Contributions

The following is a list of contributions presented in this PhD thesis. Contributions A - G are papers, and H is a technical note which is an ongoing research project. All papers and the technical note can be found in the contributions-appendix in this thesis.

A **Jonathan Dyssel Stets**, Alessandro Dal Corso, Jannik Boll Nielsen, Rasmus Ahrenkiel Lyngby, Sebastian Hoppe Nesgaard Jensen, Jakob Wilm, Mads Brix Doest, Carsten Gundlach, Eythor Runar Eiriksson, Knut Conradsen, Anders Bjorholm Dahl, Jakob Andreas Bærentzen, Jeppe Revall Frisvad, Henrik Aanæs. (2017). Scene reassembly after multimodal digitization and pipeline evaluation using photorealistic rendering. In: Applied Optics, Vol. 56, No. 27, 2017, p. 7679-7690. [SDN+17]

B Jannik Boll Nielsen, **Jonathan Dyssel Stets**, Rasmus Ahrenkiel Lyngby, Henrik Aanæs, Anders Bjorholm Dahl, Jeppe Revall Frisvad. (2017). A variational study on BRDF reconstruction in a structured light scanner. Proceedings of International Conference on Computer Vision (ICCV 2017). IEEE, 2017. p. 143-152. [NSL+17]

C Rasmus Ramsbøl Jensen, **Jonathan Dyssel Stets**, Seidi Suurmets, Jesper Clement, Henrik Aanæs. (2017). Wearable Gaze Trackers: Mapping Visual Attention in 3D. Image Analysis. Springer, 2017. p. 66-76 (Lecture Notes in Computer Science, Vol. 10269). [JSS+17]

D **Jonathan Dyssel Stets**, Yongbin Sun, Scott W. Greenwald, Wiley Corning. (2017). Visualization and labeling of point clouds in virtual reality. Proceedings of SA '17 SIGGRAPH Asia 2017. 2017. 31. [SSGC17]

E Alessandro Dal Corso, **Jonathan Dyssel Stets**, Andrea Luongo, Jannik Boll Nielsen, Jeppe Revall Frisvad, Henrik Aanæs. Virtual reality inspection and painting with measured BRDFs. (2017). Proceedings of SA '17 VR Showcase. 2017. [DSL+17]

F Henrik Aanæs, Knut Conradsen, Alessandro Dal Corso, Anders Bjorholm Dahl, Alessio Del Bue, Mads Emil Brix Doest, Jeppe Revall Frisvad, Sebastian Hoppe Nesgaard Jensen, Jannik Boll Nielsen, **Jonathan Dyssel Stets**, George Vogiatzis. (2015). Our 3D Vision Data-Sets in the Making. Abstract from The Future of Datasets in Vision 2015, Boston, United States. [ACD+15]

G Seidi Suurmets, Jesper Clement, Rasmus Ramsbøl Jensen, Jonathan Dyssel Stets, Henrik Aanæs. (2018). 3D heatmap in marketing research and marketing practice – validation of an integrating model. European Journal of Marketing. Submitted, in review.

H Technical Note: Learning Refraction with Convolutional Neural Networks. Jonathan Dyssel Stets, Zhengqin Li, Jeppe Revall Frisvad, Manmohan Chandraker.

Contents

Chapter 1

Introduction

Human-Computer Interaction (HCI) in a simplified manner is about the exchange of information or communication between a human and a computer. Humans interpret information through the use of senses, and modern technologies are enabling computers to mimic these interaction techniques. As examples of this, a speaker or a microphone utilizing text-to-speech and speech recognition, has already made the first steps for a computer to talk and understand human spoken languages. Likewise, the digital camera has enabled the computer to see, and modern screens and projectors are able to present digital images that can appear more sharp and colorful than the real world.

The title of this thesis refers to the visual interaction between the human and the computer, which covers both how the computer visually acquires and interprets the physical world, but also how to visualize digital content. The primary focus of this thesis is to target relevant challenges of visual HCI using a set of cases mainly within the field of Computer Vision but with strong ties to Computer Graphics. Figure 1.1 shows a diagram to illustrate how visual human-computer interaction is defined in this thesis.

The ultimate goal of HCI is to provide a tool for humans to better interact with a computer or machine that can assist with a specific task. Thus it is necessary to outline what the human visual system is capable of to fully understand the potential of a visual computer system.

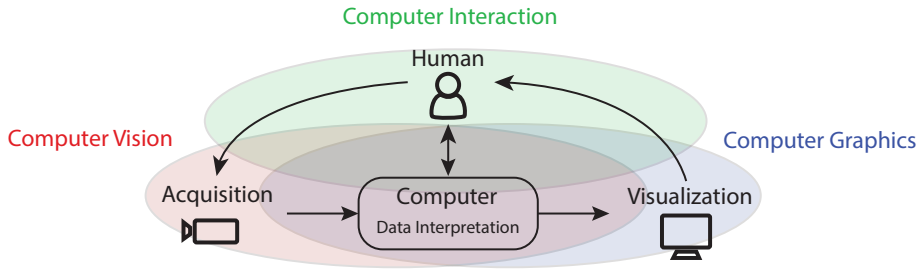


Figure 1.1: Diagram of the Visual Human-Computer Interaction pipeline as it is defined in this thesis. Three elements are surrounding the computer in the middle: The visual input to the computer here illustrated with a camera, the visual output of the computer here illustrated with a monitor, and finally the human interacting with the computer. The three central and interlinked themes are the field of Computer Vision, Computer Graphics and Computer Interaction. All three are important blocks of HCI, and sometimes goes under the common term Visual Computing.

The human visual system is a complex mechanism, and it has a significant impact on how we can interpret and navigate in the physical world. We can detect and distinguish colors and operate in a high dynamic range of light. The fact that we have two eyes adds additional valuable information so that we can see what is near and what is far away, enabling us to interpret structure and 3D information. In our early stage of life, we gain experience on how to interpret the physical world, and we learn to recognize things that we have seen before. Not only can we recognize what an object is, but we can also by appearance recognize its approximate physical properties, i.e., mass, hardness, if it is liquid or solid, rough or smooth, if it feels cold or warm, and so on. This makes the human visual system extraordinarily sophisticated and a very powerful tool for understanding our surroundings.

Because we as humans are so good at interpreting the physical world, it can, on the other hand, be difficult to convince us that something generated by a computer is real, which means that realistic visualizations are difficult to achieve.

Due to the strengths and weaknesses of the human visual system, there is a massive potential in visual computing concerning HCI. Already today advanced algorithms and technology have made many complex vision tasks possible, and some of them can outperform humans. But there are certainly still tasks that prove to be particularly challenging for today's computer vision systems.

The specific challenges are exemplified through a set of cases which is the foundation of the research projects presented in this thesis. In the next section, the relevance and applications of these cases are motivated followed by the thesis scope, objective and research methodology. Finally, a thesis outline is provided to give the reader an overview of the chapters.

1.1 Applications and Motivation

Virtual and augmented reality are two visual-based interactive techniques that prove a number of benefits with regard to HCI. They have the ability to visualize complex geometry and appearance data in a realistic and life-like way, that enhance intuitive interaction. Both of these systems rely heavily on Computer Vision to measure and interpret the environment, i.e., the ability to see, and on the ability to render content through Computer Graphics. The following cases exemplify how this has been investigated in the thesis, and while they are different in nature, they all perfectly align with the parts of the visual HCI pipeline described in Figure 1.1.

Case 1: Modeling and evaluating heterogeneous scenes with refractive geometry

A problem that is not yet solved in the field of Computer Vision is the ability to cope with refractive, or transparent objects. It is however very relevant to address this issue, as such objects can cause problems for vision systems, and they frequently appear in real-life settings. It would be suitable for an Augmented Reality (AR) system or an autonomous robot that rely on vision, to label or segment refractive objects.

Glass itself is almost invisible, so what one is actually seeing, is not so much the material itself, but rather distorted light from the surroundings. In other words, glass takes its appearance from its surroundings and this is for some applications a very nice feature - it can, for example, be used to redirect light in a camera lens. But this property also makes glass quite complex for vision tasks such as geometric acquisition, object detection, and appearance capture as illustrated in Figure 1.2.

When inspecting glass, we can see the light is refracted, reflected and absorbed. The difficult part is that these interactions can happen simultaneously, resulting in both the reflected light and refracted light being visible in the same spot.

Figure 1.2: Rendering of three glass objects. The overall shape can be interpreted, but only because of the context.

Humans can use the light interactions to interpret the shape of a glass object, but this task is very complicated for many computer vision methods because most visible light implementations are designed for light to bounce off and not pass through an object.

It is desired to acquire and evaluate the geometry of heterogeneous scenes, i.e. scenes containing objects with different radiometric properties, including glass. A method for assessing such acquisition can be accomplished by capturing all geometric components of the scene, reassemble the geometry and render the final result as a digital image. Then a comparison can be made of the rendering and the actual scene. Due to the complex nature of glass, it is a perfect fit for evaluating how accurately the geometry is captured, as small errors in the geometry will completely change the appearance of the glass object. Additionally, it is desired to investigate how analysis by synthesis can be utilized to improve the acquired geometry and appearance.

The potential use case of such a method is the ability to scan glass for error detection in a production line or scan real glass objects for catalogue renderings. Additionally, verified realistic renderings could be used to produce synthetic training data for, e.g., machine learning algorithms that learn from images of the real world. Note that glass is used to exemplify the case, but other similar complex radiometric objects could be relevant.

Case 2: Automated gaze mapping for mobile eye tracking

Eye tracking, or gaze tracking, is an important technology within the field of HCI as it allows for both interaction and control using one's eyes, but also for analyzing the usability of various HCI applications. Gaze trackers are getting more common, and can now be found in some laptops, smartphones, and Virtual Reality (VR) and AR headsets. Often the term mobile eye tracker, or gaze tracker, is used when the device is attached to the user. In this thesis, they are defined as a head-mounted pair of glasses that records what a respondent is looking at, using a traditional video camera (the technology is further described in Section 2.3.1). This means that we only know what the respondent is looking at relative to a video stream, and not to the actual physical scene. Utilizing Computer Vision techniques we wish to design a method for automated gaze mapping between the user and the physical scene for mobile gaze trackers. The specific use case in this regard is described subsequently.

The company iMotions provides a platform for synchronizing data from biometric sensors such as gaze trackers, Electroencephalography (EEG) and Galvanic Skin Response (GSR). This case focuses specifically on mobile gaze trackers, which are used in many different fields of research to study human behavior in various scenarios. When conducting a study with gaze tracking, there is often a goal of monitoring if a respondent is looking at certain key objects or regions in a given scene. These regions are known as Areas of Interest (AOI), and have to be manually annotated by a human for the software to determine if and in which area, a gaze point falls within. Figure 1.3 shows an example of how an AOI can look like.

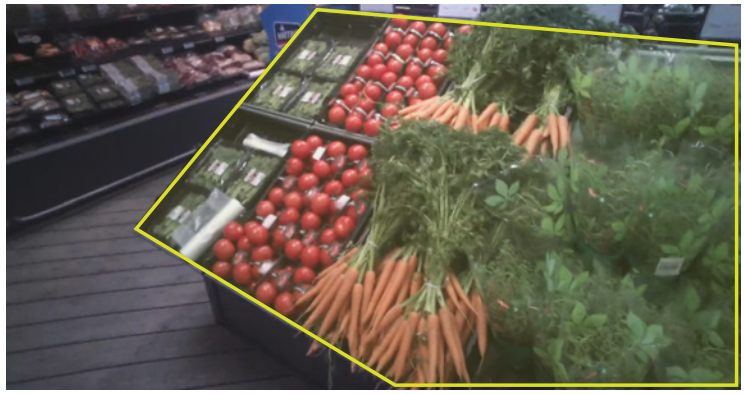


Figure 1.3: An AOI (inside the yellow border) shown on top of a video frame from the gaze tracker.

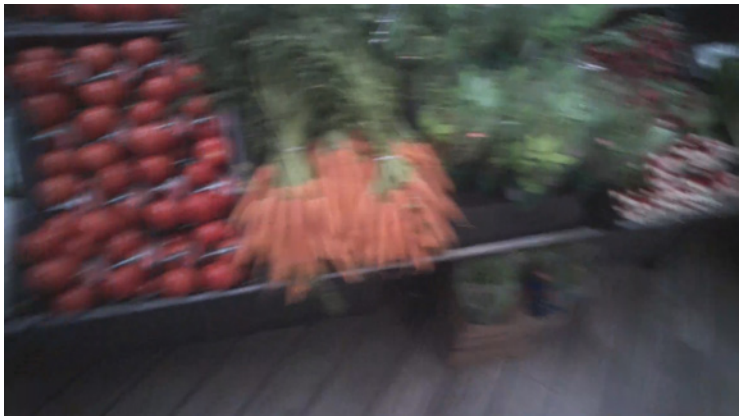


Figure 1.4: Video frame from the gaze tracker: Motion blur often occur in the video due to rapid head movements in conjunction with poor lighting conditions.

At the beginning of this PhD project, such a task was performed manually by drawing a polygon with mouse-clicks in every frame of the gaze tracker video stream. Recordings with mobile gaze trackers can last from seconds to several hours allowing for extensive studies of where the respondent looks and focuses on. This results in millions of mouse-clicks which is cumbersome, error-prone, and a very time-consuming process. A way to ease this process is to utilize the video feed from the mobile gaze tracker to create a semi-automatic computer-assisted annotation method. With the AOI annotated, it is possible to map gaze points and construct an attention heat map.

The method should require a minimum effort from the operator and preferably only need commercial off-the-shelf hardware components. Alongside these constraints, a challenge of this case is the video output from the mobile gaze trackers, which can be noisy at times, as shown in Figure 1.4.

Case 3: Virtual reality as an interactive tool for geometry and appearance visualization.

VR is a technology that has existed for many years but has gained quite a lot of attention recently due to improvements in both VR hardware and the evolution of the Graphics Processing Unit (GPU). It is now something that potentially soon could be found in almost every home and to some extent, it practically already is, as any smartphone can be converted into a simple VR device. A

full spherical view and a one-to-one physical control can be achieved with VR technology, enabling the possibility for a lifelike and intuitive experience. Thus it is an excellent candidate to be used for visualization and interaction of data for HCI applications, and the potential use-cases are many. VR can be used for robot control, simulations for training and education, entertainment and gaming applications, and inspection of complex data.

Accordingly, it is desired to explore the potential of VR as a tool for visualizing complex geometry and appearance data. Currently, a vast amount of VR applications belong to the entertainment industry, and the computer-generated content is still kept at a minimum for performance reasons and necessary high frame rate requirements. But it is expected that more complex geometry and appearance models can be visualized as time progress, and it is worth investigating what types of applications are suitable for VR.

The three above-mentioned cases are specifically targeted in the contributions presented in this thesis and will be further addressed in the contributions section.

1.2 Scope and Objective

The work in the thesis addresses several aspects of the visual HCI pipeline with the focus on the process of accurately going from the physical to the digital domain, and back in the form of a graphical visualization. While this is a broad topic, the scope of the research presented in this thesis, aims to investigate and propose solutions to specific tasks, which contribute to optimizing accurate acquisition and realistic visualization.

The spectrum of the research projects presented in this thesis contributes to solving a set of challenges exemplified in the cases described in the previous section. From these challenges, a set of four central topics has been selected to group the contributions. The objectives of the topics are presented below, and the links between the topics in relation to HCI are illustrated in Figure 1.5

Geometry and Appearance Digitization: Scenes in the real world are often heterogeneous with respect to radiometric properties, where diffuse, transparent, specular or translucent objects can be present simultaneously. The objective is to target multimodal acquisition of heterogeneous scenes, specifically including glass, and acquire geometry and appearance data to be used for accurate digitization as motivated in Case 1. Furthermore, it is investigated how analysis by synthesis can be utilized to evaluate acquisition techniques using quantitative

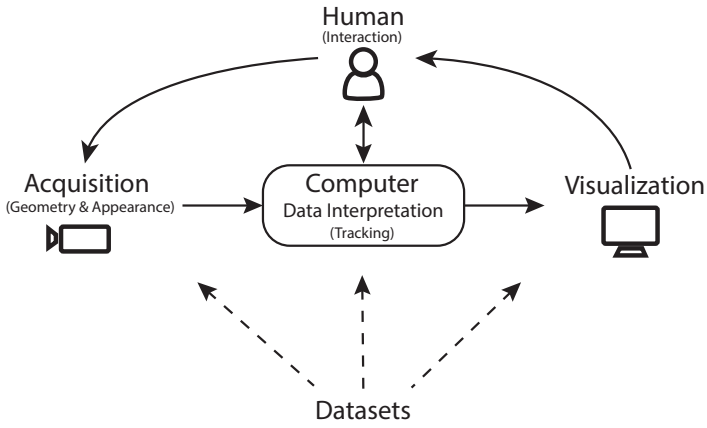


Figure 1.5: Illustration of the links between the thesis topics with basis in Figure 1.1. The different parts of the HCI pipeline is explored in the contributions of this thesis. The approach is data driven, therefore datasets play an important role in all steps of the pipeline.

comparison. This has many practical applications for assessing realistic renderings and 3D acquisition methods used in both research and industry context.

Tracking: The objective is to provide a method for AOI tracking and gaze mapping specifically for Case 2, contributing with a technical implementation for industrial purposes. Besides applying to the mentioned case, image- and video-based tracking, in general, are highly relevant for AR applications and navigation of autonomous vehicles and robots.

Visualization and Interaction: The objective is to target visualization methods specifically for geometry and appearance digitized from physical objects. It is explored how VR can be used as a method to visualize complex geometry and appearance data, as mentioned in Case 3, and how gaze data from eye trackers can be augmented onto geometry for accurate and intuitive data interpretation. Specifically, AR and VR applications are highly relevant as modern computers and hand-held electronics can perform fast computations with vast amounts of data, enabling the opportunity for fluid real-time interfacing.

Datasets: The objective is to provide data to support the above mentioned topics. This includes a multimodal dataset containing geometry and appearance data obtained using different acquisition methods, a gaze dataset with images, video, and gaze data from eye-tracking experiments and finally a synthetically generated dataset to be used for, e.g., learning algorithms and assessment. The

collected and generated data is a crucial component for designing and evaluating the methods presented in this thesis.

This defines the scope of this thesis, and the objectives of the contributions to the four topics. Based on these, the methodology is presented in the following section.

1.3 Methodology

The work presented in this thesis aims meet the above defined objectives mainly using a practical and data-driven approach. This means that much work has gone into collecting data and designing frameworks, dictated mainly by the cases introduced previously in this chapter. While such a process is intricate and time-consuming, it also allows to genuinely design the data and framework to match the problem we are trying to solve.

Both acquisition of real world data and generation of synthetic data has been carried out for the work presented in this thesis. The experimental hardware equipment includes a high precision industrial robot with an interchangeable mount, an industrial CT scanner, a structured light scanner, a mobile gaze tracker, industrial cameras and RGB-depth sensors. Synthetic data has been generated with physically based rendering tools and GPUs. The methodological contributions presented in this thesis stem primarily from the field of Computer Vision with an overlap of Computer Graphics. Hence, it is a multidisciplinary effort that includes research areas such as image-based tracking, 3D vision, deep learning, Physically Based Rendering (PBR), VR, AR and gaze-tracking.

1.4 Thesis Outline

To present the work of this PhD, the thesis is divided into the chapters outlined below. After the introduction chapter, the background theory is introduced followed by a summary of contributions and a conclusion of the thesis. A list of papers can be found on page ix. The chapters contains the following material:

Background and Related work - chapter 2 introduces a selected set of the fundamental concepts and theories used in the contributions of this thesis. Image acquisition, camera geometry, and calibration is introduced in the beginning

of the chapter, followed by a range of geometry acquisition methods. Then follows an introduction to graphics related concepts, including how to understand and simulate light interaction with materials and how to realistically synthesize images with regard to color and appearance. Finally, there is an introduction to interaction techniques which describes VR and AR technologies and methods and how to interpret and visualize data from gaze trackers. Note that the background and related work chapter is not directly a one-to-one correspondence with how the contributions are presented in Chapter 3. Instead, the background and related work are presented so that the first part of the chapter is Computer Vision related topics, the second part is Computer Graphics related topics, and the final part is interactive technologies as demonstrated in Figure 1.1.

Contributions - chapter 3 describes the contributions of this thesis, which has been ordered in the following scheme according to the thesis objectives: Geometry and Appearance Digitization, Tracking, Visualization and Interaction, and Datasets.

Conclusion - chapter 4 concludes the contributions of this thesis.

Contributions - appendix contains the publications mentioned throughout this thesis. They are included in their submitted format. A technical note has been added to this thesis describing an ongoing research project related to the presented contributions.

Chapter 2

Background and Related Work

This chapter provides an overview of the concepts and technologies used for the research presented in this thesis. Due to the broad scope of this thesis, there will only be a brief introduction to relevant topics, however related work is referenced for further and more detailed descriptions on the topics. The individual topics are divided into the following three central themes aligned with Figure 1.1, namely:

Vision and Acquisition which describes how to acquire and digitize geometry from the physical world. The section contains an introduction to basic camera geometry and various geometric acquisition methods.

Graphics and Visualization which describes how to render images, obtain realistic appearance, and interpret light and colors.

Data Interpretation and Interaction describing the methods for interaction and how data is interpreted and used for visualization. This includes an introduction to hardware components such as gaze trackers, VR and AR technologies.

2.1 Vision and Acquisition

In the field of computer vision the camera is commonly used to acquire geometry from a given scene. This thesis focuses mainly on 3D vision, which, as opposed to 2D imaging, enables the ability to get structural information from our physical three-dimensional world.

Consequently, this section will outline the basic theory of how the physical world is digitally modelled using a camera. Accurate acquisition and reconstruction of geometry have a wide range of applications such as the preservation of cultural heritage [GPBS14] and additive manufacturing [GRS10, ZDES16]. Equally important is the general geometric interpretation of the physical world, which is used in, e.g., autonomous vehicle navigation [JGBG17]. So the following will describe the basic camera geometry and calibration procedure followed by 3D acquisition techniques such as structured light scanning and Structure from Motion (SfM). Additionally, the hardware tools used for geometry capture is introduced in this section.

2.1.1 Camera Geometry

To acquire the world in 3D from a vision system, it is necessary to understand camera geometry and how a camera depicts the world. A simplified way to describe a camera is by the pinhole camera model [HZ03], which works the following way. A box with an infinitely small hole, also known as a pinhole or aperture, is facing a scene illuminated by an arbitrary light source. The light, or photons from the light source, hit the scene and bounce off the scene elements in all sorts of directions. The light, or photon paths, is referred to as rays or light rays. As the light rays are bouncing off the scene, most of them will continue and hit other objects, while some of them will continue through the pinhole. Since the pinhole is infinitely small, it can be considered a single point, so all rays that enter the box will intersect this point. The result is a flipped 2D projection of the illuminated scene on the backside of the box, which is denoted the image plane. The distance from the pinhole, or aperture, to the image plane is named the focal length. For a more intuitive understanding of the projection, the non-flipped frontal image plane is defined as being one focal length in front of the aperture. Figure 2.1 shows an illustration of the pinhole camera model, where an illuminated 3D scene is projected onto the 2D image plane through the aperture.

In a real digital camera, the image plane consists of a 2D grid of sensors that converts light waves into electrical signals. The two main technologies used for

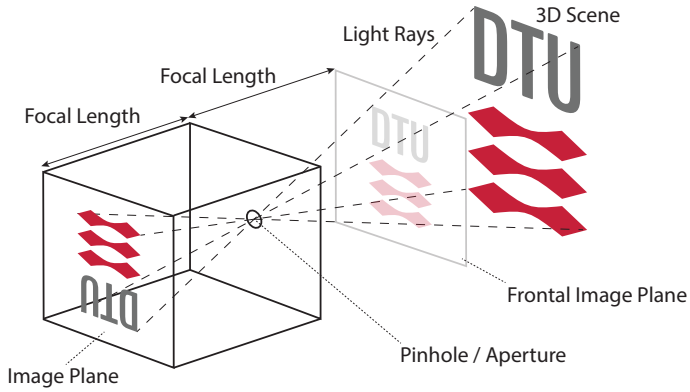


Figure 2.1: Illustration of the pinhole camera model. The light from an illuminated 3D scene is projected through the aperture into the backside of the camera box revealing a 2D projection.

this is either Charge-Coupled Device (CCD) or Complementary Metal Oxide on Silicon (CMOS), which each has various and proven benefits such as power consumption and sensitivity, however, the CMOS technology is most commonly used in digital cameras today [Sze11]. For both technologies, it holds, that the photon energy is integrated over a time period, exposure time, and converted to an electrical signal for each sensor in the grid. This results in a 2D discrete matrix of values also referred to as pixels.

The pinhole camera model can be used to mathematically describe the geometric relationship between a 2D image and the 3D scene as illustrated in Figure 2.2. Two coordinate systems are defined: the three-dimensional world coordinate system, often measured in meters, and the two-dimensional image coordinate system measured in pixels.

The image point corresponding to a world point is found where the line between the world point and the focal point intersects the image plane. After a point is projected to the image plane, that pixel alone does no longer contain any information about the relative depth of its 3D position.

The relation between the world point Q to the image point q is a geometric transformation defined as the camera matrix $\mathbf{P}$ [HZ03]:

$$q = \mathbf{P}Q, \tag{2.1}$$

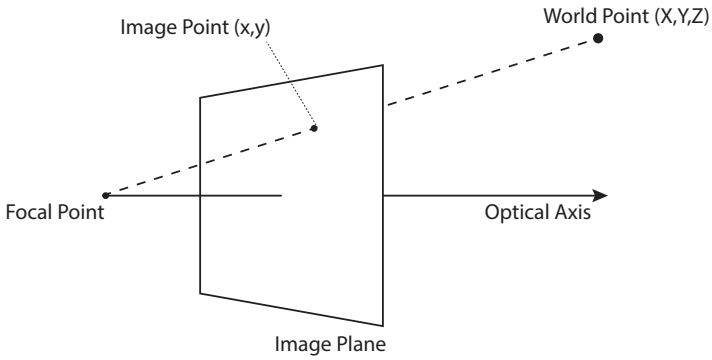


Figure 2.2: Relationship between a 3D world point and its corresponding image point.

where

$$\mathbf{P} = \mathbf{A} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}. \tag{2.2}$$

The rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ represents the extrinsic parameters, and $\mathbf{A}$ is the intrinsic matrix:

$$\mathbf{A} = \begin{bmatrix} f & s & \Delta x \\ 0 & f & \Delta y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.3}$$

with the focal length f, the coordinates of the optical axis $[\Delta x \ \Delta y]$ and the skew s. The skew parameter can be neglected and set to zero in most cases for a modern camera [HZ03]. These parameters are estimated when performing camera calibration, which is further described in Section 2.1.2. A real camera is equipped with a lens system, and so it should be noted that radial distortion can be present and should be accounted for in the pinhole camera model if high accuracy is needed. The radial distortion occurs as a deformation in the image referred to as a 'barrel'- and 'fisheye' distortion. This happens because the incoming light rays are bent more towards the edges of the lens compared to the optical center. The distortion can be approximated with the following

expressions [HS97, Sze11]:

$$\begin{aligned} x' &= x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ y' &= y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6), \end{aligned} \tag{2.4}$$

where (x', y') are the corrections of the coordinates (x, y). (k_1, k_2, k_3) are the radial distortion coefficients of the lens and $r = \sqrt{x^2 + y^2}$ is the distance, or radius, from the optical axis.

If the lens and the camera sensor is not perfectly parallel, then tangential distortion should also be accounted for in the camera model. The tangential distortion is modeled by [HS97]

$$\begin{aligned} x' &= x + [2p_1 xy + p_2(r^2 + 2x^2)] \\ y' &= y + [p_1(r^2 + 2y^2) + 2p_2 xy], \end{aligned} \tag{2.5}$$

where (x', y') are the corrections of the coordinates (x, y) and (p_1, p_2) are the tangential distortion parameters.

The above theory is crucial when designing a computer vision system, but equally important is estimating and tuning the parameters from the mentioned equations. This is done using camera calibration and is described next.

2.1.2 Camera Calibration and Pose Estimation

To accurately model the camera, a number of distortion parameters should be accounted for. A common method to perform a camera calibration is presented in [Zha00] and only requires the camera to observe a planar pattern from minimum two different positions. A checkerboard pattern with known dimensions is a good calibration object, since corners, or saddle points, are easy to detect with high precision. Points on the planar pattern are detected in multiple images, and knowing the relative distance between these points, it is possible to estimate both the intrinsic camera parameters, the distortion coefficients, and the pose transformation between the camera positions, i.e., the extrinsic camera parameters.

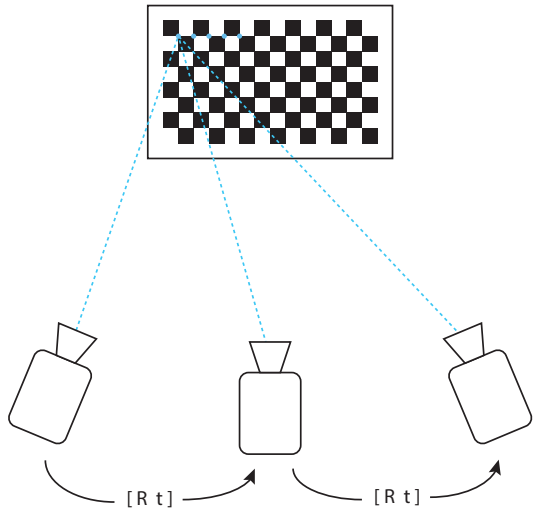


Figure 2.3: Camera calibration: A checkerboard is observed from three different camera positions, and saddle points are detected on the checkerboard (blue dots). A relative camera position can be calculated for each view, knowing corresponding points across each view.

Figure 2.3 demonstrates a typical calibration setup, where points on the checkerboard are detected from three different camera positions. Note that either camera or checkerboard should be moved, but all calibration points need to remain visible in the images. The result of the calibration can be verified by re-projecting the detected points, using the estimated parameters, onto the images of the checkerboard. The offset between actual points and re-projected points is referred to as the re-projection error. The method is implemented in MATLAB [Mat, Bou] and OpenCV [Bra00]. More images can increase the precision of the camera calibration, and one should ensure that calibration points are present in all areas of the image. However, too many calibration images can cause overfitting and result in worse calibration results, which can be monitored in the re-projection error.

Pose Estimation or extrinsic calibration can likewise be obtained from a set of 2D projected points from, e.g., a known 3D object. Techniques to accomplish this is described in [Sze11] where the relative transformation between an object and the camera is derived. When estimating the pose for calibration purposes, a checkerboard calibration target as described above is typically used [Mat, Bou], but self-calibration methods exist [FLM92, Har93]. Image feature points, or other types of known markers, can also be used for this purpose.

Stereo Camera Calibration is the process of determining the transformation $[\mathbf{R}\ \mathbf{t}]$ between the two cameras, and follows the same principle as above. Before estimating the relative position between the stereo pair, each camera should be calibrated individually following the method described above. For the stereo calibration, it can be beneficial to allow the calibration algorithm to iteratively update the camera parameters using minimization while estimating poses [HZ03]. It should be noted that the same type of checkerboard pictures can be used for calibrating both the stereo pair and the individual cameras as long as all checkerboard points are visible in all images from both cameras. The re-projection error is again used for validation and optimization and is obtained by projecting points from one stereo camera to the other.

Working with a calibrated camera system is advantageous in many computer vision applications. In the next subsections, two techniques of obtaining geometry with a camera system are presented; Structured Light Scanning where a calibrated stereo camera setup is beneficial, and SfM where pose estimation proves to be useful.

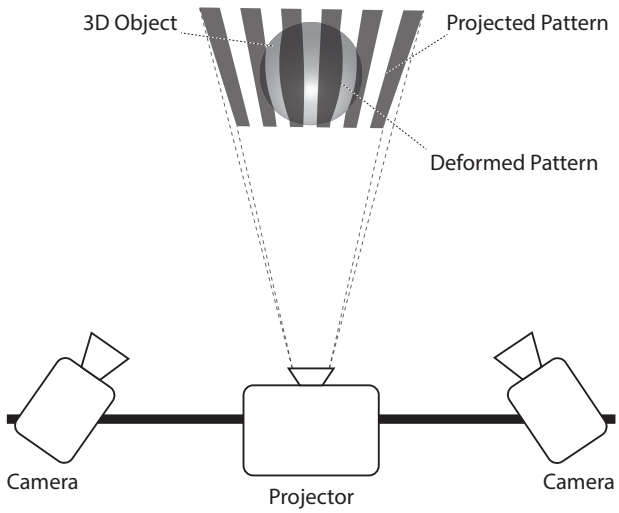


Figure 2.4: Two cameras and a projector constitute a structured light scanner. A coded pattern is projected onto a 3D scene, and the deformation of the projected pattern reveals the geometry of the scene.

2.1.3 Structured Light Scanning

A structured light scanner is a device that can perform a 3D scanning of a scene or object using one or more calibrated cameras and a light projector. The scanner can be realized from rather inexpensive hardware components, yet it can produce high-quality point clouds if the system is properly set up, as presented in [MT12, EWPA16]. It works by projecting a known structured pattern onto the surface of a scene, which is then decoded in the images from the cameras. Figure 2.4 sketches the structured light setup used for research presented in this thesis. It consists of a calibrated stereo camera setup and a projector, all three devices mounted on the same baseline.

A number of structured light patterns are compared in [SFPL10], and binary patterns and Gray coding is selected for our system due to its robustness and simplicity. A set of patterns comprised of unique vertical light stripes are projected onto the static scene. The cameras capture images for each projected pattern, and the intersection between a scan line and a projected stripe is used to compute 3D points by triangulation [Gen11].

Compared to other structured light methods, the binary coding technique allows a pixel to only have two values, and it is, therefore, less sensitive to surface char-

acteristics [Gen11]. But the general limitation of the structured light scanner is the ability to require geometry with diffuse radiometric properties. Because the system relies on the presence of a light pattern on the scanned object, the surroundings should ideally be dark for optimal results. In addition, the presence of specularity, transparency, or significant absorption will contaminate the performance. This complicates digitization of many real-life scenarios as such scenes are often heterogeneous and therefore hard to acquire only with structured light.

Our implementation is not speed optimized, so it is crucial for the scanner rig and scanned scene to remain static while scanning. Real-time implementations exist [Wil16], but are not needed for our applications. It should be noted that points can only be obtained from the area lit by the projector, thus, the effective output from a single scan will be a 2.5-dimensional reconstruction. However, multiple scans can be acquired from different angles, either by moving the structured light setup [SDA11], or by rotating the scanned object [EWPA16]. Point clouds can subsequently be combined for more coverage of the object using pose estimation information and refined with an Iterative Closest Point (ICP) algorithm as proposed in [LHZB+16].

Despite its disadvantages, a structured light system is an excellent tool for diffuse surfaces, and therefore it is used for the acquisition of such geometry in the research presented later in this thesis.

2.1.4 Structure From Motion

Structure from Motion (SfM) [KVD91] is a technique to obtain 3D geometry and a camera motion path from a sequence of 2D images. It is important to introduce this concept of this method, as it is essential for the contribution to Case 1. SfM has the advantage of only relying on a single camera, and unlike structured light, it is not relying on a projected pattern, while it still remains a non-contact method. To create a correspondence between images, SfM uses image features such as corners [HS88], or more advanced features like the Scale-Invariant Feature Transform (SIFT) [Low04] or Speeded Up Robust Features (SURF) [BETVG08]. Feature points are matched by their descriptors across all images, and Random Sample Consensus (RANSAC) [FB81] is applied to filter out false positive matches, leaving only inlier points within a given threshold [HZ03]. The inlier matches are used to determine extrinsic camera parameters, and the result can be refined with a bundle adjustment [TMHF99]. Figure 2.5 shows a camera capturing images of a 3D scene from two different positions. Feature points are detected on each image in the sequence and matched pairwise, as illustrated in Figure 2.6 (left). The points are used to sparsely reconstruct the scene and estimate the camera poses, as illustrated in Figure 2.6 (right).

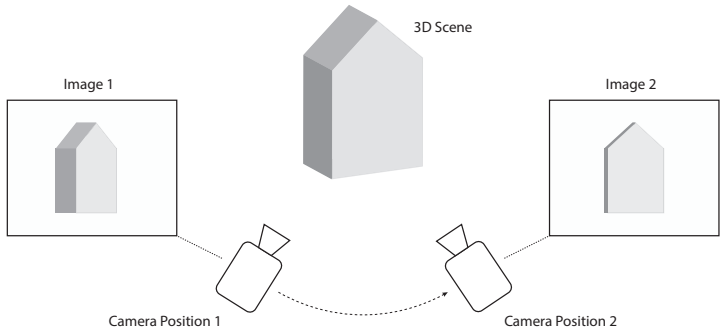


Figure 2.5: A 3D scene observed from two different positions and the corresponding 2D images from each view.

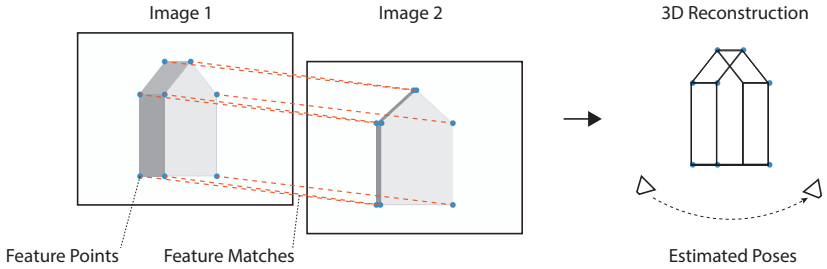


Figure 2.6: Detecting feature points in each image (blue dots) and matching them across images (red lines). The information is used to estimate a 3D reconstruction of the feature points and camera poses.

It should be noted that the camera used for SfM in this thesis is calibrated to compensate for distortion in the acquired images. The result is often a sparse point cloud, and more dense results can be obtained using the method described in [FP10].

Because the method relies on image features, the quality of the images has a significant impact on how well it performs, both with respect to image resolution and image distortion. It is likewise essential that image features exist in the image, and that these are repeated across the sequence of images. This means that very smooth or soft convex scenes with a homogeneous appearance should be expected to have a lack of features and therefore can be hard to reconstruct. On the other hand, a scene with a repeating pattern, i.e., repeating features, may result in confusing the matching of true positives. Finally, scenes containing elements whose appearance changes drastically based on view direction,

such as glass or chrome, is a significant problem for a feature-based approach. However, the SfM technique is desirable with respect to hardware requirements and portability. The accuracy is very dependent on the restrictions mentioned above, and will often be far from a calibrated scanner rig, like, e.g., structured light. Consequently, this method is well suited for applications without a dedicated scanner system, mobile AR and 3D reconstructions of large-scale data [WBG+12].

This concludes the methods, utilized in this thesis, for obtaining geometry using a camera system. In the following subsection, another technique is introduced to obtain geometric information of physical objects without the information from visible light.

2.1.5 Computed Tomography Scanning

A Computed Tomography (CT) Scanner uses X-rays to produce tomographic images of the scanned object. Using an X-ray source and X-ray detectors, 2D slices are generated based on the attenuation of the media which the X-ray beam is traveling through. The slice can be considered a traditional 2D image, comprised of pixels as 2D squares in the XY plane. The thickness of the slice is the Z axis, comprising a volume element, or voxel, with the pixels in the XY plane [Rom95].

The CT slice image is reconstructed from many independent detector measurements. Errors in these measurements may lead to unwanted artifacts in the reconstructed image, and some of these are identified in [BK04]. An example is streak artifact in the reconstructed image, occurring due to the beam hardening effect, where the mean energy of the X-ray beam changes when traveling through neighbouring materials with high-density variation [BK04, SDN+17]. Avoidance or handling of such artifacts should be kept in mind when designing a system relying on CT scans, to avoid losing accuracy.

The CT is advantageous because objects can be scanned all the way through, revealing hidden geometry, and transparent and semi-transparent materials also appear on a CT scan. However, the physical construction size of a CT scanner is often large and not easily portable. See Figure 2.7 for an image of the scanner used for the research presented in this thesis.

Figure 2.7: Images of the CT scanner utilized for research presented in this thesis. The left photo gives an idea about the physical size of the machine, and the right photo shows the inside of the scanner.

2.1.6 Imaging Robot

Much of the image data used for research in this thesis is generated with an imaging robot [AJV+16]. The robot is a 6-axis industrial robot arm, which can repeat pre-programmed motion paths with high precision [AD15]. Two cameras and a projector are mounted on the robot arm, so both stereo images and structured light scans can be captured. A photograph of the robot mount is shown in Figure 2.8. The robot is located in an enclosure that blocks almost all incoming light, and both robot and enclosure are covered with diffuse black paint to eliminate light bounces. Furthermore, programmable LED light sources are mounted inside the enclosure, enabling to control both the amount of light and light direction. The setup is well suited for experiments, where many images or structured light scans need to be captured within a controlled environment. The robot provides data for pose estimation, but this information can also be acquired by calibrating the extrinsic parameters of the mounted stereo camera system.

A series of vision based techniques have been presented in this section, and all of these are used in contributions presented later in this thesis. The next section provides an equivalent overview of the graphics and visualization tools.

2.2 Graphics and Visualization

It has previously been described how to acquire geometry using different vision techniques. The following sections focus on the Computer Graphics related top-

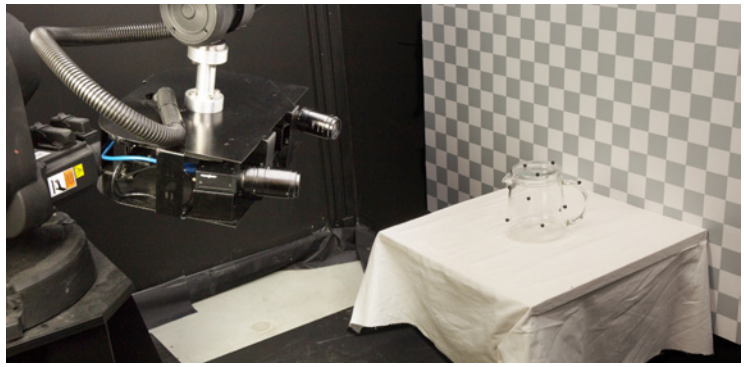


Figure 2.8: Photo of the imaging robot. The mount contains two cameras and a projector to be used for stereo images and structured light. The picture is from Contribution A.

ics, which explains how to generate images of acquired data and how light and appearance is typically handled by a computer. This includes how to simulate images using ray tracing, how to obtain realistic global illumination with High Dynamic Range (HDR) images, and finally, how to evaluate appearance by perception. The techniques presented in this chapter are important, as they are used in many applications in, e.g., movie production, computer games, advertisement, additive manufacturing, but also in, e.g., geometry from appearance research applications. In this thesis, the techniques are used to evaluate geometric acquisition using analysis by synthesis, for producing realistic VR content, and to synthesize training data for machine learning.

2.2.1 Light Simulation and Realistic Rendering

While geometry acquisition from images is a way to digitize the physical world from visual appearance, rendering can be understood as the reverse action, namely, visualizing digital data as it would appear, if it were the physical world. This section gives a brief understanding of the basic technique used by computer software to describe the appearance of a digital model. First, consider the pinhole camera model shown in Figure 2.1, where the 3D world is projected into a 2D image. The same principle yields for ray tracing [App68], but now a ray is traced from the focal point or the pinhole. The focal point is denoted the eye, and now only the frontal image plane is considered. When rendering the image, the color of each pixel should be determined by what light eventually hits that pixel. This task is realized by tracing rays in the scene [Gla89], i.e., ray tracing. Figure 2.9 shows how pixels are colored by tracing rays in a 3D Scene.

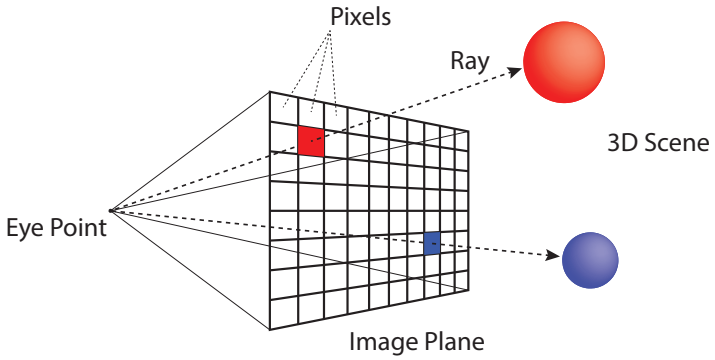


Figure 2.9: Ray tracing of two pixels in a simple scene with a red and a blue 3D object. Rays are traced in the direction from the eye point to the pixel point on the image plane. The pixel's color is determined by the appearance of the object it hits.

Light interacts differently depending on the type of surface it hits, so only tracing a single ray per pixel is not sufficient to produce an accurate simulation. Recursion is needed, so the result from multiple rays are combined for each pixel, for effects like refraction, reflection and shadows to appear realistically [Whi80].

On a very basic level, the light interaction in a medium can be characterized as reflection, transmission and absorption. Depending on the properties of the material, there will be a combination of these basic interactions that all contribute to the appearance of the material. Figure 2.10 illustrates the ray interacting between two media. Note that the absorbed ray is not reflected, nor refracted, but transformed into another type of energy such as, e.g., heat. The type of media and the surface boundary, on a micro-scale level, between the two media, will determine how light interact and thus the appearance. More surface roughness will typically result in a more diffuse appearance.

Rays can reflect and refract both from one medium to another and internally in one medium, which can cause effects such as subsurface scattering and specular highlights. The color of the object is determined by the ray interaction, specifically which wavelengths are absorbed and reflected respectively.

To synthesize the ray interaction with materials, there exist mathematical models to describe how rays should interact depending on its incidence angle. An example of such a mathematical model is the Bidirectional Reflectance Distribution Function (BRDF) [Nic65]. PBR follows the laws of mathematics and

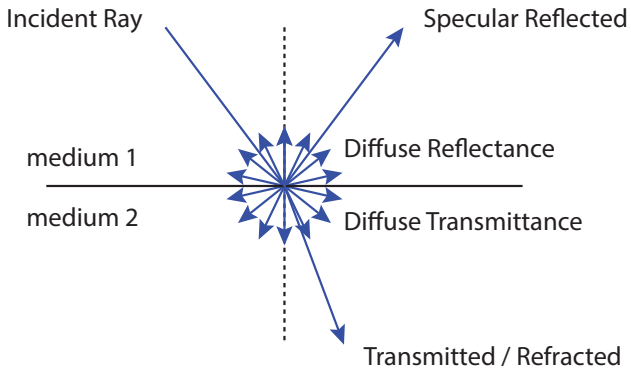


Figure 2.10: Different types of ray interactions between two media. An incident ray travels in medium 1 and interacts with the boundary to medium 2. From here it can reflect or refract, and depending on the properties of the boundary and the types of media, the ray can scatter in multiple directions causing either a diffuse or specular appearance.

physics [PJH16], so images produced with PBR will appear realistically as if it was an image captured of a real-world scene. This is an important link to HCI, as PBR is the key to visualize geometry and materials in a way that the human visual system is familiar with. PBR is a complex task because humans very often can tell, perhaps not right away, that a computer-generated image is synthetic. The following PBR tools have been evaluated during this project: Cycles [Prob], Mitsuba Physically Based Renderer [Jak], and NVIDIA® OptiX™Ray Tracing Engine [PBD+10]. The OptiX framework provides ray tracing on the GPU and proves to be well suited due to its speed performance, realistic results, and the fact that it is highly customizable.

An essential factor in reproducing scenes as realistic renderings are to accurately model illumination. Next, it is explained how images can be used for global illumination of a scene, which is a favorable method to achieve a realistic appearance.

2.2.2 Image-Based Lighting

In an outdoor daylight scenario, the sun will be the only dominant source of light, and in an indoor scene, the light will most likely originate from lamps. It is important to remember that indirect illumination typically constitutes a vast

amount of the light that hits an object in a real-world setting. In an outdoor daylight scenario, this is, e.g., all the light that originates from the sun, but has bounced off from various surfaces there may be. Using global illumination is a favorable method to synthesize realistic appearance and examples of this are presented in [Deb12].

The method in [Deb08] proposes to acquire a HDR spherical image of the surroundings to be used as a global illumination source as opposed to specific synthetic light sources. In this image-based lighting method, each pixel in the spherical image contributes to the scene as a light source, and these are assumed to be very far away from the scene.

Methods for acquiring environment maps involves photographing a mirror sphere [Deb08] or using a rotating camera at a fixed location. When photographing a mirror sphere, there will be an almost complete coverage of the scene in the reflection from the sphere. This method requires a warping of the image from the reflection of the sphere [RHD+10] and the method specifically used for this thesis is described in [SDN+17, Nie16]. The other method of using images from a rotating camera requires the images to be warped and stitched to a sphere. Using stitched images can create a high-resolution environment map, but the image acquisition is slower than the mirror sphere, as multiple angles need to be captured depending on the cameras field of view. Almost the full environment map is captured from only a single angle when using a mirror sphere, but it will result in an uneven resolution, as a large area is covered by few pixels close to the edges of the mirror sphere, and the camera will be visible in the reflection. Finally, there is also the option of using a 360-degree camera.

The effect of HDR and how to interpret the data is explained next.

2.2.3 High Dynamic Range Imaging and Tone Mapping

High Dynamic Range Imaging (HDRI) is a technique to capture images in a broader dynamic range of illumination. The advantage is that a scene containing both low and high illuminated areas is captured more accurately, and unwanted effects known from traditional photography, such as over- and under-exposure, is avoided. The illumination from sunlight is in the order of $10^5 cd/m^2$, while indoor lighting is $10^2 cd/m^2$, and moonlight can be as low as $10^{-1} cd/m^2$ [RHD+10], i.e., a difference of 6 orders of magnitude. The human visual system is excellent at adapting light sensitivity in different situations, so we can see during the day and night, giving us the ability to operate within a highly effective range of illumination.

As opposed to the human visual system, traditional cameras are using 8-bit per pixel for respectively the red, green, and the blue color channel to store information. In other words, a pixel can distinguish 256 intensities for each color channel, resulting in $256^3 = 16.7$ million colors per pixel in total. In some cases, this information is sufficient for capturing a scene, but in many cases, information is omitted. This is visible as regions of an image being either over- or underexposed. Light sources, or the area lit directly by a light source, will be saturated in order for darker areas, for example, illuminated by indirect illumination to be within range and vice versa. When there is a need for modeling realistic illumination, to be used for realistic visualization purposes, it is crucial to have the HDR information [Deb08].

The method for acquiring HDR images of the real world from a standard camera is to capture multiple exposures of the same scene and then combine the information in the images [RHD+10, DM97]. Using this method, the scene is sampled in different ranges, so both low and high intensities are captured relatively and finally stored in commonly 16- or 32-bit.

HDR displays exist but are certainly not common property. Most displays today have 8-bits per color channel, so a transformation is needed in order to display an HDR image on a Low Dynamic Range (LDR) display. One of the first appearances of tone mapping for computer graphics is presented in [TR93]. Here, the importance of sensation preservation when presenting images on a screen is highlighted for gray-scale images. It is noted that good tone reproduction is dependent on radiance, luminance, and brightness. Several Tone Mapping Operators (TMO) exist [RSSF02, Ash02, RHD+10], and [EMU17] contains a study of TMO for HDR video.

The above-mentioned methods are utilized in the contributions of this thesis and the actual ray tracing implementation is beyond the scope of this thesis. The following describes how to evaluate the accuracy of the rendered images using color difference.

2.2.4 Color Difference

A measure of color difference can be used as a score to tell how different or how much alike two colors are. A difference can be understood in numerous ways, but in this thesis, the desired parameter for comparison is the human perceptual difference. A very basic color difference in the RGB space could be the l_2-norm,

which for two RGB colors is calculated in the following way:

$$d_{RGB} = \sqrt{(R_2 - R_1)^2 + (G_2 - G_1)^2 + (B_2 - B_1)^2}. \tag{2.6}$$

However, this measure does not take the human visual perception into account. It is known that the human visual system is not equally capable of seeing the differences in respectively reds, greens, and blues. So a shift in the green color channel compared to the numerically same shift in the blue color will by a human perceptual difference measure not necessarily be the same for the two cases.

A distance metric, ΔE^*_{ab}, is introduced, which operates in the CIELAB color space. CIELAB is an approximation to a uniform color space designed for perceptual color difference [HRV97]. It is a non-linear re-mapping of the XYZ color space to compensate for the non-linearity perception in the human visual system [Sze11]. The difference between two $L^*a^*b^*$ colors is given by

$$\Delta E^*_{ab} = \sqrt{(L^*_2 - L^*_1)^2 + (a^*_2 - a^*_1)^2 + (b^*_2 - b^*_1)^2}. \tag{2.7}$$

A Just-Noticeable Difference (JND) or visual threshold value for the ΔE^*_{ab} exists, which in [RHD+10] is defined as the minimum amount of increment that an observer can distinguish. $\Delta E^*_{ab} \approx 1$ is defined as JND [HRV97] meaning that two $L^*a^*b^*$ colors, whose distance is lower than this value, is said to be indistinguishable for the average human.

Due to inaccuracies, improvements has been made to eq. 2.7, and it was updated in 1994 to ΔE^*_{94} [HRV97]. This was once again improved in 2000 to become the CIEDE2000 color-difference formula [LCR01, SWD05], which is the most recent color difference equation. Here, the difference between two $L^*a^*b^*$ colors is given by

$$\Delta E^*_{00} = \sqrt{\left(\frac{\Delta L'}{k_L s_L}\right)^2 + \left(\frac{\Delta C'}{k_C s_C}\right)^2 + \left(\frac{\Delta H'}{k_H s_H}\right)^2 + R_T \frac{\Delta C'}{k_C S_C} \frac{\Delta H'}{k_H S_H}}, \tag{2.8}$$

with $\Delta L'$ being the lightness difference, $\Delta C'$ the chroma difference, $\Delta H'$ the hue difference, and S_L, S_C and S_H being their respectable compensation functions. k_L, k_C and k_H are the parametric weighting factors and usually set to 1, but varies depending on the application. R_T is the rotation term to compensate for performance in the blue hue region.

The methods introduced in this section described how a digital scene can be realistically rendered into an image and how that image can be perceptually compared to another image. This is useful when pixel-wise comparing a rendered image to a real photo of a given scene to determine their similarity. The next, and final, section of this chapter will introduce the interaction technologies used in the contributions of this thesis.

2.3 Data Interpretation and Interaction

The following sections will describe the hardware and interaction tools that are mentioned and examined in this thesis. Furthermore, there will be a description of how to interpret and visualize specifically data from a gaze tracker.

2.3.1 Gaze Tracking

Gaze tracking, or eye tracking, is the process of measuring in what direction a person is looking, or what a person is looking at. This is interesting in a broad spectrum of applications and is found in product analysis, marketing research, psychological studies of human behavior but is also relevant for HCI related research [PB06].

In this thesis, gaze tracking has been used as a tool for mapping human attention to a scene, and the Tobii Pro Glasses 2 has been used for gaze tracking. The focus has therefore not been on research in gaze tracking, but merely on analyzing the output from the gaze tracking to build 3D heat maps to visualize the attention of a respondent. To understand current state-of-the-art gaze tracking, there will in the following be a brief introduction to the development of this technology.

In [Duc07] multiple eye-tracking techniques, both intrusive and non-intrusive, are addressed. Early methods consist of measuring muscle activity around the eye (electro-oculography) or attaching traceable contact lenses to track eye movements with search coils. Video-based methods have however turned out to be the most widely used today, and this method is implemented in most modern eye trackers. The basic concept is to have a camera monitoring a respondent's eye and detecting the relative pupil position. Often, Infrared (IR) light is emitted onto the eyeball, and the corneal reflection of the IR light relative to the pupil will improve the precision of the gaze tracking. For further technical details, a survey of video-based eye tracking techniques is presented in [HJ10].

Figure 2.11: 'Tobii Pro Glasses 2' gaze trackers. The image is from [Proa].

One can think of two types of gaze tracker devices; screen based gaze trackers and mobile gaze trackers. The screen-based gaze tracker is a device that is often attached and calibrated to a display, monitoring only gaze coordinates in that specific screen space. The advantage here is that it can be directly controlled what is viewed on the screen; thus it is simple to autonomously record exactly what a respondent is looking at. The downside is that stimuli are constrained to be viewed on the screen, and the respondent has to be within the range of the gaze tracking device. The mobile gaze trackers, on the other hand, are head mounted directly in front of the respondent's eye. This enables the possibility of tracking gaze in real life scenarios, but this also contributes to more complex post-processing of the data.

The Tobii Pro Glasses 2 [Proa], from now on referred to as Tobii Glasses, shown in Figure 2.11, have been used for gaze tracking related research in this thesis. They are a head-mounted pair of glasses equipped with a forward-facing camera that can record what is being looked at. IR emitters and sensors are on the inside of the glasses, facing the respondents' eyes, and used to calculate the gaze direction. The glasses have to be calibrated, but afterwards, they can provide synchronized gaze data in pixel coordinates of the video stream from the forward facing camera. Additionally, the Tobii Glasses are also equipped with an Inertial Measurement Unit (IMU), more specifically gyroscope and accelerometer, enabling more accurate motion estimates [KS11].

A general challenge when working with mobile gaze trackers is the complexity of post-processing the video data. It contains no information about what is viewed, so a method for tracking AOI is needed to avoid too much manual annotation as described in Case 2. It should also be mentioned that the output video of the gaze trackers can be noisy. This is most likely a combination of a person easily making swift head movement in conjunction with a relatively small image sensor and lens system. Note that experiments and research regarding gaze tracking

in this thesis are recordings of indoor scenes with limited light.

To interpret the data from an eye tracker, it is often favorable to visualize both what a person has been looking at, while simultaneously showing the gaze data in a way that match the experiment. Next subsection will introduce a method for visualizing gaze data.

2.3.2 Heat Maps for Gaze Visualization

There are many ways to interpret gaze data, consequently there exist a number of methods for visualizing such data. Some studies are determined to use all gaze points from the eye tracker, i.e., the raw data, while other studies prefer to use only fixations, i.e., an aggregation of gaze points based on a time or displacement span with a given threshold. These metrics are further described in [HNA+11]. The method one should use for visualizing gaze data depends on the study and the type of gaze tracker. An overview of visualization methods is provided in [TKM+], and an attention map, or heat map, turns out to be a popular method, but should be used correctly as discussed in [Boj09].

A heat map is a visualization technique showing data as colors, often as an overlay on another set of data. A common way to visualize gaze data is to augment a heat map on top of an image of a given AOI, and discussion of heat maps as a visualization method for such data is provided in [pM07]. For this thesis a red, yellow, green color coding is used for visualization, where red denotes the highest value, fading to yellow, fading to green being the lowest value. The heat map provides an intuitive way to interpret 2D gaze points from one or more respondents. Figure 2.12 shows an image of cereal shelves in a supermarket with a heat map overlay generated from gaze data.

There is no standard for generating heat maps, but suggestions for kernel shapes is presented in [pM07]. Consequently, it should be assumed that a heat map is used for visualization purposes only unless it is explicitly known how the heat map is constructed. In this thesis, a 2-dimensional Gaussian kernel is used for generating the color-coded heat maps

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \tag{2.9}$$

where x, y are the points in the kernel, i.e., pixels coordinates when augmenting the kernel onto an image. σ is the variance of the kernel and for gaze analysis purposes, this could be the certainty of the measured gaze point, if this measure

Figure 2.12: An image with augmented heat map generated with the iMotions Software [iMo]. The figure is from contribution C.

is available. If depth information is available, the kernel size could be dependent on the distance from the respondent's eyes to the gaze point on the scene. In other words, the top point of the kernel is placed in the gaze or fixation coordinate, and the width of the kernel is determined by how precise that point is, or how much it should cover. Kernels for all gaze or fixation points are combined to produce the final heat map.

Visualizing heat maps is relatively simple for gaze data recorded with stationary gaze trackers of a static scene. For temporal data, such as video sequences or data from mobile gaze trackers, the visualization becomes more complex as the respondent is now looking at a dynamic scene. Different solutions for this are surveyed in [TKM+].

Case 2 presents the problem of analyzing and visualizing data from mobile gaze trackers. Here we have dynamic content from the gaze tracker video-stream which is determined by the head position and movement of the respondent. A method for visualizing gaze data for dynamic content is presented in [KW13]. However, it is relevant to consider a 3D heat map visualization as opposed to the traditional 2D visualization and several works addresses this opportunity. In [SNDM11, MHB14] attention maps are added to the surface of virtual 3D geometry, [Pfe12] proposes 3D attention volumes, and a 3D saliency map using a SLAM based approach is presented in [PSF+13]. This related work is highly

relevant to the problem described in Case 2 which will be further elaborated in the contributions chapter.

It should finally be noted that the video-stream from the gaze tracker along with gaze coordinates is used in this thesis. The IMU sensor data provided by the gaze tracker would be relevant to include, but it is beyond the scope of this thesis.

2.3.3 Virtual and Augmented Reality

The final part of this chapter provides an introduction to Virtual and Augmented Reality, as these technologies play an important role in this thesis as mentioned in Case 3. VR is a technology to display virtual content, commonly through a head-mounted display, and can give the user a sense of being submerged in a virtual world, while AR is a technology to combine or overlay virtual content on top of the physical world in, e.g., real-time. In this thesis, VR is demonstrated for displaying geometry and appearance measured on real-world objects, while AR is referenced as an application for some of the presented contributions. Consequently, these technologies are briefly introduced in the following.

Virtual Reality is a technology that enables full visual submersion into a computer-generated scene by covering a large part of the user's field of view commonly with a digital display. Modern VR solutions consist of a headset and often hand-held controllers that act as the interface with the virtual world. VR requires a high frame rate and high-resolution images for each eye to give a fluid and realistic experience, so the computation power needed to generate interactive VR content is often extensive. However, this problem seems to be solved by the modern GPU to an extent where it is acceptable. Current popular VR systems includes the HTC Vive [HTC], Occulus Rift[Ocu], and Playstation VR[Pla], but also a smartphone can be transformed into a VR headset.

Augmented Reality can be in the shape of many things, such as a screen-based application overlaying content on top of a camera feed, a computer-generated projection onto physical objects or a semi-transparent display. AR is a well-suited technology for HCI, and new AR technologies continuously appear. An example is the head-mounted Microsoft Hololens [Mic], which provides AR holograms through a total internal reflection display technique, and also maintains mapping and localization. In many cases, AR strongly relies on a well-performing vision system that can handle real-time processing of image data.

Localization and Tracking

For many VR and AR technologies, tracking or localization in relation to the physical world is a requirement. An IMU can be used to track head movement, and additional sensors can be used to compensate for the degrees of freedom. Frequently, the terms outside-in versus inside-out tracking is mentioned in VR and AR systems.

Outside-in tracking means that one or more observers are placed in the surrounding environment to observe the user. An example could be an estimation of the pose of the human body like the Microsoft Kinect, or trackable markers attached to the body or device like the Playstation VR [Pla] and the Oculus Rift [Ocu]. The latter technology currently proves to be the most accurate, and more external sensors can be added to improve accuracy. It generally also requires less or no electronics to be attached to the device or body. However, it is an easy victim for occlusion, i.e., something is blocking the path between sensor and marker. Also tracking is restricted to a limited area defined by the field of view of the sensor.

Inside-out tracking means that one or more tracking sensors, such as a camera, a 3D scanner, or similar, is attached to the device or user. Localization is achieved by observing stationary trackable features on the surroundings, such as markers, image features, structured light patterns, or similar. The displacement of these features can be used for pose estimation. The advantage is that the tracking system is not restricted by a limited area, and is often self-contained, so no additional setup is necessary. However, such a system can be bulkier as more electronics, and computational power is needed on the device. The Microsoft Hololens [Mic], for example, is utilizing the inside-out tracking and is entirely self-contained.

Use cases

Both VR and AR are, despite some psychological issues [SS01], tools well suited for HCI, since complex data can be made easily understandable for humans even without much prior computer experience or similar prerequisites. The overlay benefit from AR seems to be a good option for information and control applications, such as machine operation and manufacturing as exemplified in [EPF+17, WBE+13, NOCM12]. VR is well suited for simulations, where one has to be fully immersed such as training for scenarios that are difficult to recreate in real life, or controlling machinery that one cannot be physically nearby. Additionally, VR is an obvious way to visualize three-dimensional data

for direct inspection and interaction, and can potentially make such interaction more intuitive than on a conventional two-dimensional screen.

This chapter provided a brief overview of the research themes presented in this thesis, namely: vision techniques to capture physical objects, graphics techniques to realistically render and visualize a digital model and a selected set of human interaction technologies used for the digital visual domain.

Chapter 3

Contributions

This chapter provides a summary of the contributions and perspectives of the research projects. On an abstract level, the contributions consist of three main themes, namely how to deal with refractive objects, photorealistic rendering for computer graphics applications, and systems for advanced and realistic complex applications for HCI. The arguments for addressing these three themes are as follows:

Refractive objects are a part of our everyday environment, so to obtain full-fledged 3D reconstructions outside the laboratory environment, it is crucial to be able to handle such objects. This has proven to be a complex task in vision [BeKN13, INKPAL+13], and the contributions in this thesis provide a step in the direction towards a better understanding by proposing methods to handle refractive geometry with an experimental and data-driven approach.

For computer graphics to produce photorealistic content, there is a need for actually comparing renderings to real photos. An early approach to this was presented in [GTGB84] and later addressed in [UWP06], but recently this has somewhat fallen in the background, as renderings are often not quantitatively compared. Contributions in this thesis have made an attempt to revitalize this part of computer graphics by reinitiating quantitative assessment using state of the art methods in 3D capturing and rendering.

In the commercial use-case, Case 2, it is illustrated that there is a need for complex HCI by mean of wearable devices. In general, there is a trend that AR and VR technologies are used more widely than before in HCI applications such as education, training, data monitoring, machine operation, navigation, and entertainment. These systems are complicated to work with; consequently, use cases are needed to understand them better and how to design them, which is part of the contributions of this thesis.

This section will outline some of the more specific contributions and relate them to the above. To set the contributions of this thesis in perspective, the main results have been ordered according to a loose and overlapping four-element taxonomy: geometry and appearance digitization, tracking, visualization and interaction, and datasets.

The titles of the contributions are as follows, and this list can also be found on page ix:

A Scene reassembly after multimodal digitization and pipeline evaluation using photorealistic rendering. [SDN+17]

B A variational study on BRDF reconstruction in a structured light scanner. [NSL+17]

C Wearable Gaze Trackers: Mapping Visual Attention in 3D. [JSS+17]

D Visualization and labeling of point clouds in virtual reality. [SSGC17]

E Virtual reality inspection and painting with measured BRDFs. [DSL+17]

F Our 3D Vision Data-Sets in the Making. [ACD+15]

G 3D heatmap in marketing research and marketing practice - validation of an integrating model.

H Technical Note: Learning Refraction with Convolutional Neural Networks.

3.1 Geometry and Appearance Digitization

Much research relating to geometry and appearance digitization has been conducted previously, and many problems have been solved for a wide variety of methods. But it remains a challenge how to do a complete acquisition of geometry and appearance from a scene containing objects with diverse radiometric properties.

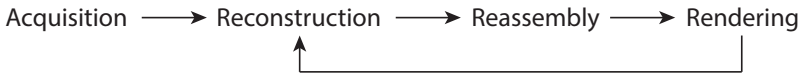


Figure 3.1: Main steps of the pipeline from contribution A. First, data is acquired using suitable appearance and geometry acquisition methods. Then the data is reconstructed and reassembled into the same coordinate space to finally produce a photorealistic rendering of the scene. The rendering in the final step is quantitatively compared to a reference image, and the last three steps can be repeated for refinement if necessary.

The work of this thesis relating to geometry and appearance digitization contributes to the understanding of refractive objects and provides a systematic method for empirical evaluation of computer graphics as described in Case 1.

Contribution A, address a multimodal scene acquisition method, which can capture, reconstruct and realistically render a scene with objects of both refractive and diffuse radiometric properties. An end-to-end digitization pipeline has been realized that can perform the following tasks: acquire both geometry and appearance using suitable acquisition methods, reconstruct the data, and then reassemble all individual components. Finally, the combined data is rendered and evaluated. Thus, it is possible to improve or estimate parameters using analysis by synthesis, by repeating steps in the pipeline and re-evaluate the final result. Figure 3.1 shows the outline of the main blocks in the pipeline.

The pipeline utilizes existing methods and techniques known from the fields of computer vision and computer graphics, but despite that, the implementation proves many practical challenges. The proof of concept setup consists of a glass object on top of a cloth and a checkerboard patterned backdrop. A photo of the setup compared to the final realistic rendering of the scene is presented in Figure 3.2.

The glass is scanned with a CT scanner, and the remaining scene is scanned with a structured light system. A marker-based approach is used to accurately combine the different elements in the reassembly step of the pipeline. Figure 3.3 shows the principle of the marker-based approach. The markers are detected in both types of scans, which prepare the ground for an accurate transformation between reference frames.

The radiometric properties of the diffuse part of the scene cannot be captured with our structured light system, so these are instead captured as BRDFs for each type of material subsequently. To match appearance and geometry, a

(a) Rendered Image **(b)** Reference Image

Figure 3.2: Rendered image produced by the pipeline from Contribution A, compared with the reference photo.

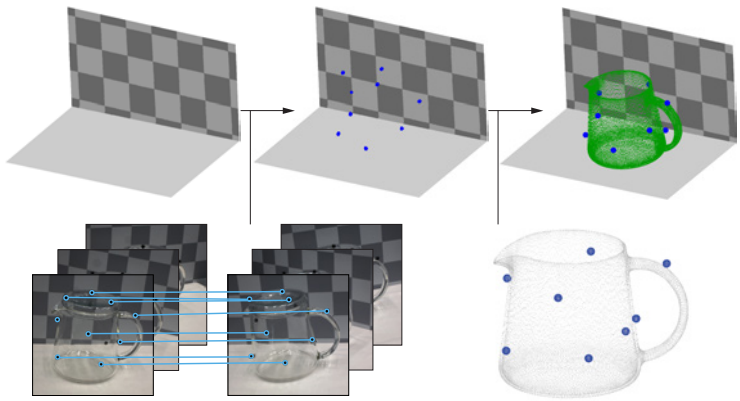


Figure 3.3: Marker-based approach to combine geometry from two different acquisitions methods: Markers are detected in images of the scene and triangulated. Markers are also detected on the CT reconstruction of the glass object, and this relation is used to compute a geometric transformation. The figure is from contribution A.

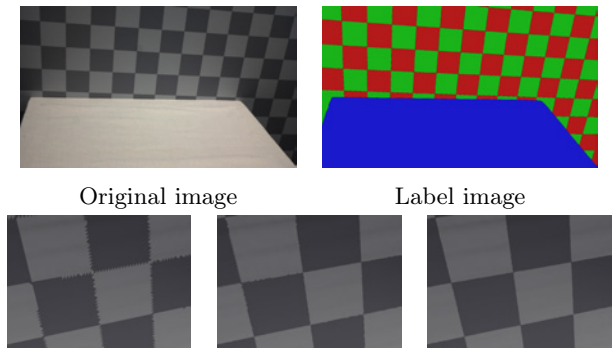


Figure 3.4: Segmentation and labeling of the geometry: The first row shows how the original image is segmented, resulting in a label image. The second row shows the effect of the micro-polygons, where sawtooth boundary artifacts are suppressed from left to right. The figures are from contribution A.

micropolygon labeling approach is used to assign the measured BRDFs to the correct place in the geometry [SDN+17]. Figure 3.4 shows how images are used as a basis for segmentation, which is then projected onto the geometry. To get rid of unwanted boundary artifacts, the mesh is subdivided.

The output of the pipeline is a realistically rendered image of the captured geometry, see Figure 3.2, and is comparable on a pixel to pixel basis with the photographed scene. This enables the ability to perform analysis by synthesis, i.e., the ability to tune parameters and directly observe and compare the changes in the final result. The comparison of improvements is presented in Figure 3.5. In summary, contribution A allows for quantitative assessment of appearance and geometry in an acquired multimodal scene. To the best of our knowledge, the ability to make empirical comparison of transparent geometry and directly quantify photorealism of such a scene has not been done previously [SDN+17].

In contribution A, appearance was captured subsequently from the geometry acquisition process. Ideally, these two steps should be fused as the surface geometry holds valuable information about appearance and vice versa.

In contribution B it is investigated whether appearance and geometry can be captured directly in a structured light scanner. The study uses a data-driven BRDF reconstruction approach to investigate the impact of uncertainties of different parameters [NSL+17]. Four pre-measured BRDFs with varying levels of specularity are used for the experiment. They are rendered individually on a sphere positioned in a scene with an added light source to simulate a desired

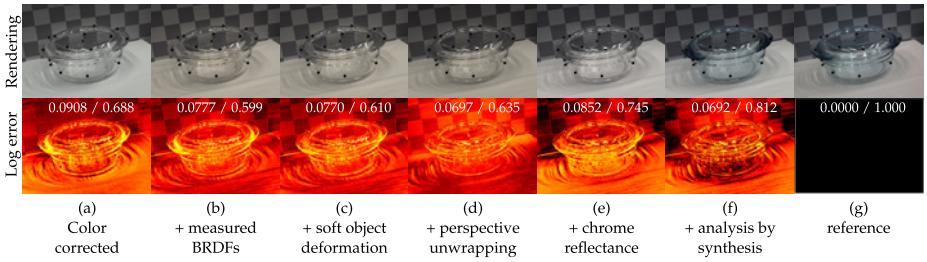


Figure 3.5: Comparison and evaluation of improvement steps in the pipeline to the reference image (g). The result improves from left to right, and the scores on the images are respectively root-mean-square-error (high to low) and structural similarity index (low to high). This demonstrates how each step is directly evaluated and compared. The figure is from contribution A, where color map and similarity measures are further explained.

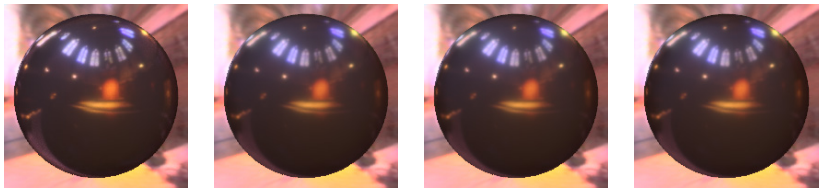


Figure 3.6: Perceptual difference is used as a metric to evaluate the accuracy of the reconstructed reflectance. The far left image is a rendering of the reference BRDF, while the next three are renderings of the BRDF reconstructions from varying the light intensity. The renderings are from contribution B.

structured light setup. Noise is introduced to the vertex positions, vertex normals, light position, and light intensity, and the BRDFs are reconstructed from the synthetic images and re-rendered on a sphere with global illumination (as described in Section 2.2.2). The re-rendered BRDFs are compared alongside the original BRDFs, and the images are tone mapped and perceptually compared using the ΔE_{00} color difference measure (as described in Section 2.2.4). An example of the perceptual compared images are shown in Figure 3.6. A table with the perceptual impact of the noise parameters is provided in contribution B and contributes to design considerations towards building a fused appearance and geometry capturing system.

Contribution H presents a learning-based approach for obtaining geometry from refractive objects. Inspired by [EPF14] who predicts depth from single images, we train a Convolutional Neural Network (CNN) with a dataset of rendered

images of refractive objects for depth estimation. Utilizing computer rendering enables easy access to large sets of data within a reasonable time frame. The network is trained with images of both diffuse objects and refractive objects. The results are promising, and ideally, the network can also work on real images. This can potentially be a method for rough estimation of refractive geometry.

In conclusion, the work described above contributes to systematic methods for acquiring appearance and geometry of specific scenes with refractive objects. It also provides a method for empirical evaluation of computer graphics. Consequently, this is a step towards obtaining more precise geometry of refractive objects and more realistic computer visualizations.

3.2 Tracking

Tracking is crucial in applications utilizing navigation or positioning in a real-world setting. Examples of this include AR, autonomous vehicles, and navigating robots, and here there is often a need for real-time tracking systems. In other cases, post-processing of data is necessary, and an off-line tracking system is sufficient. Case 2 states the need for tracking AOI in video data from mobile eye trackers. The task is however linked with a series of challenges originating from the mobile gaze tracker as described in Section 2.3.1.

Contribution C suggests to capture the AOI with a consumer Digital Single-Lens Reflex (DSLR) camera. With that data, it is possible to reconstruct a 3D model of the scene and build an image based feature descriptor database for reference. The 3D model is of sufficient quality for visualizing the AOI, which could not have been achieved using the gaze tracker alone. The visualization is further addressed in Section 3.3.

Utilizing the feature descriptor database it is possible to match features from the gaze tracker video stream and determine its relative position. This information is valuable as it enables to track exactly what a respondent is looking at, which is also the main problem described in Case 2. Figure 3.7 shows the reconstructed 3D point cloud of the AOI and the relative position and look direction of the mobile gaze trackers, while Figure 3.8 shows the 3D AOI projected onto a single video frame from the gaze trackers.

The pipeline contributes to an end-to-end system, with a direct industrial application, but is also applicable for other HCI applications within, e.g., AR. In conclusion, this method allows noisy optical data to be used as a basis for optical localization and tracking with the help from higher quality reference data.

Figure 3.7: 3D reconstruction of a shelf with cereal boxes. The colored markers ranging from red to yellow is the estimated poses of the gaze tracker attached to a respondents head. The figure is from contribution C.

Figure 3.8: Back projection of the 3D AOI (shown in color) into a frame from the gaze tracker video (shown in grayscale). Here, it is illustrated how it is possible to find correspondence between the 3D AOI and the gaze data, despite a challenging image quality from the gaze tracker. The orange dot and lines show the gaze point and gaze path respectively. The figure is from contribution C.

Figure 3.9: Gazepoints projected onto the 3D AOI as 3D attention heat maps, shown for two different scenes: A cereal shelf and a wine shelf. The figure is from contribution C.

The method has made the foundation for the tracking solution implemented in the software provided by the industrial collaborator of this PhD project.

3.3 Visualization and Interaction

This section addresses the research related to visualization and interaction. The primary focus here is Case 3 and the visualization part of Case 2 and two VR applications for visualizing geometry and appearance.

Contribution C was mentioned in the previous chapter as a method for tracking AOI. But this contribution also suggests a novel approach to visualizing heat maps in 3D as opposed to the traditional 2D visualization. As mentioned in Section 3.2, the relative pose information of the mobile gaze trackers is known, and consequently, it is possible to project the gaze points directly onto the 3D model. A heat map can be replicated in 3D and overlayed directly on the AOI model using a Gaussian kernel applied at the gaze points (as described in Section 2.3.2). Figure 3.9 shows two AOI with heat map overlays and estimated poses of the mobile gaze tracker.

A mobile gaze tracker is mostly used to analyze real-world 3D scenarios, so it seems obvious and intuitive to visualize the gaze data in 3D. Furthermore, this method of mapping gaze points also has a series of benefits as opposed to traditional 2D heat map visualization [JSS+17]. The 3D heat maps combined with information about the relative spatial position of the gaze trackers is extremely valuable information when analyzing gaze data and is used practically in contribution G.

AR would be an obvious application for this and would be rather uncomplicated since both pose and the 3D information is available. One could imagine going to the location of the AOI and inspect the heat map directly on the physical scene through either a tablet or a wearable AR device (see Section 2.3.3). VR would also be a useful tool for visualization and would possibly require more comprehensive data. If so, one could be fully emerged in the data without physically being present at the actual location. This is exactly what has been investigated in Contribution D.

Contribution D is an application for displaying point clouds or sequences of points clouds in VR while also enabling the user to interact with the data. That way, it allows for direct human interaction with many data points while maintaining three-dimensional visual feedback. Using a depth sensor, it is possible to visualize 3D data acquired from the real world in real time. Figure 3.10 shows a screenshot from the VR application. The method can be used to visualize and inspect point cloud data as, for example, the data generated in Contribution C, or obtained by other 3D point cloud acquisition techniques. The advantage here is that one can see 3D data in a lifelike fashion without actually being physically present. Another strength of VR point cloud data interaction is to have a fast and intuitive method to label room scale datasets. The labeled data can also be used for data-driven algorithms for classification and segmentation purposes in, e.g., the domain of autonomous robots, or to edit and label real-world animated VR content.

Another important part of visualization concerning HCI is the ability to represent appearance realistically in an interactive environment. This is addressed in the VR painting application in Contribution E.

Contribution E presents a demo-application for visualizing BRDFs directly on 3D models in Virtual Reality. A BRDF, models light based on surrounding illumination, so VR is an excellent tool for an intuitive way to represent realistic appearance in a realistic environment. The application enables a selection of BRDFs to be applied with a paintbrush on the surface of a 3D model. The light interactions with these realistic material appearances can be inspected with a point light source or with global illumination from interchangeable environment maps (as described in Section 2.2.2). User testing proves that the application contributes to a fast and intuitive way to inspect BRDFs and 3D models under varying illumination circumstances. The application is a novel idea that both artists and engineers working with 3D models and appearance can benefit from. Figure 3.11 shows a screenshot from the VR application.

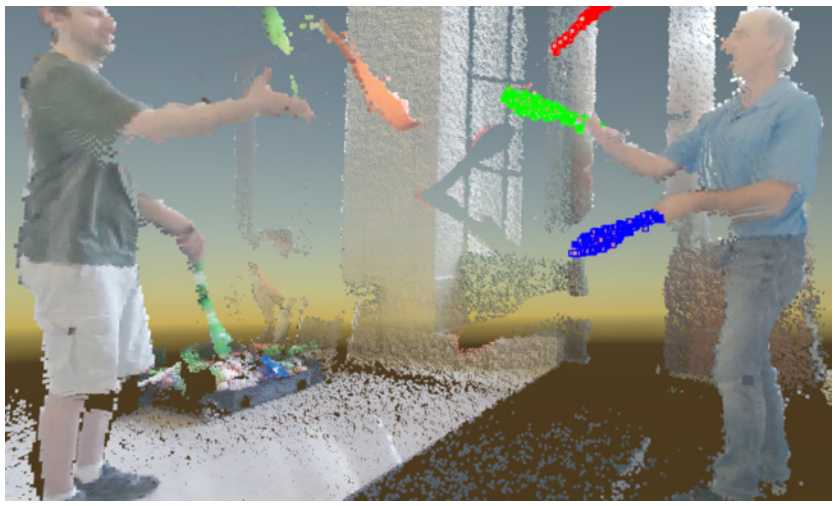


Figure 3.10: Screenshot from the VR application from Contribution D. A single frame from an animated point cloud is visualized in VR, and juggler pins on the right side are selected and labeled by the user. Labeled points in the point cloud are semi-automatically tracked in subsequent frames of the point cloud sequence.

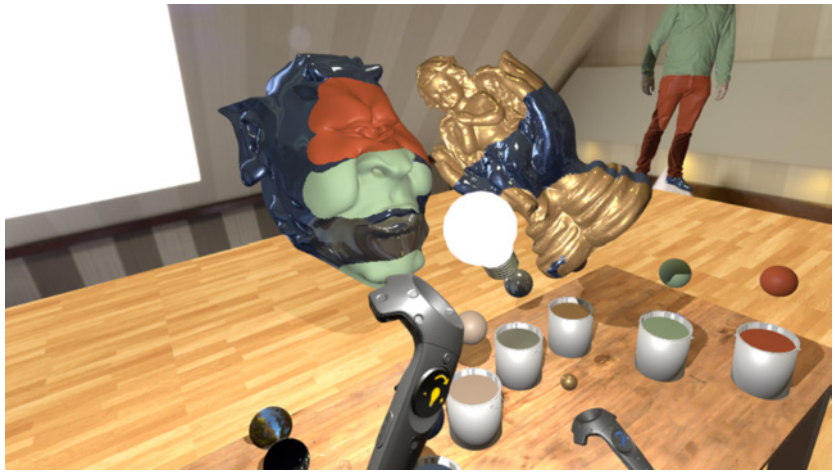


Figure 3.11: Screenshot of the VR application from contribution E. Each paint bucket on the table holds a different BRDF and can be applied to the 3D models with a paintbrush attached to the hand-held controller. Light interacting with the surface of the models is inspected by moving the light bulb or 3D model around, also with the hand-held controllers. The global scene illumination can be changed to a selection of HDR environment maps available.

In both Contribution D and Contribution E there were experiments with the use of haptic feedback when interacting with geometry. This can potentially enhance the feeling of the digital data from the virtual world being present as it was the physical world.

This concludes the contributions made within visualization and interaction techniques for HCI in this thesis. The contributions presented adds valuable results with regard to making HCI a more smooth and intuitive experience.

3.4 Datasets

This final section of the contributions chapter will briefly discuss the importance of the data that has been used for the above-mentioned work. Data is critical in many research projects within the field of Computer Vision and is often the foundation for evaluation of a method. Additionally, they can give useful insights in realistic scenarios as mentioned in Contribution F, and using dedicated hardware and software it is possible to approximate realistic settings. It is, for example, possible to simulate a motion path or control the lighting of a given scene.

As mentioned in the methodology section, the research in this thesis has been mainly data-driven. This also means it has been necessary to generate and acquire data for evaluation for most of the presented projects. To give an idea about how to obtain the data, two of the datasets from this thesis is briefly introduced in the following.

The multimodal pipeline described in Contribution A, contains a range of different acquisitions and is the most comprehensive dataset of this thesis. An overview of the captured data is listed below, and this data is also published in its raw format:

CT Scans	Volumetric data of scanned glass objects.
BRDF Materials	BRDF reconstructions of the materials used in the scene, in the MERL BRDF format [MPBM03].
Calibration Image Data	Checkerboard images for camera calibration and pose estimation. Color checker images for color calibration, and chrome sphere images for mapping of global illumination.

Scene Image Data	Multiple view stereo images of the scene with and without glass objects.
Structured Light Image Data	Images of the scene with projected structured light patterns to be used for 3D scene reconstruction.

Additionally, processed data from intermediate steps of the pipeline and final rendering data is also provided. The data is intended for other researchers to reproduce and improve the results obtained in contribution A. The data can also be used to test the performance of algorithms, where there is a need for precise geometric ground truth data of glass objects. This could be geometry estimation of refractive objects from images, detection and segmentation of glass, or undistortion of backgrounds through refractive objects.

Much of this data is captured with an industrial robot arm with high repeatability precision, but it is still important to keep track of calibrations between captures and ensure scene components does not move between captures. Programming of the robot path is also necessary, and this process was iterative to provide sufficient coverage of the scene after it was set up. The same applies for pose estimation, where the pose-calibration-target needs to be visible from all positions. Additionally, the glass objects are CT scanned, and it is essential that the objects and markers do not change between these different types of acquisitions. Due to these facts, the data acquisition process is complex and time-consuming, consequently, this data is considered a valuable contribution.

Contribution H uses a dataset of realistic renderings of refractive objects. The dataset consists of various shapes rendered on different backgrounds while providing ground-truth data such as depth-maps, normal-maps, labels, and light path information. The data has been rendered using the OptiX ray tracing engine, and the images have been realised using a vast amount of customized 3D models and environment maps. The setup of the automized renderer, model and environment map selector has required multiple iterations to ensure the output images are of the necessary quality. The dataset is intended for machine learning purposes and to explore if computer-generated data can be used as training data for geometry estimation tasks using such images.

It can be complex and time-consuming to acquire data for testing and evaluating algorithms. However, it is often worth the effort to get the actual data that is desired to be used for the specific task. The dataset from Contribution A is published, and the dataset from Contribution H is planned to be released. These are considered important contributions as they can be used as benchmarks by others to improve upon the work presented in this thesis or contribute to further development of algorithms in similar research projects.

This concludes the contributions chapter where all the research projects have been presented and grouped as the four topics: geometry and appearance acquisition, tracking, visualization and interaction, and datasets. In the next chapter, the work will be summarized and concluded.

Chapter 4

Conclusion

A series of important challenges has been addressed in this thesis, that all have in common to improve the bridge between the physical and digital domain. The challenges have been exemplified through a set of cases that overlap the research areas of Computer Vision, Computer Graphics and Computer Interaction. Due to the multidisciplinary nature of the work in this thesis, the contributions were grouped into the four topics; geometry and appearance digitization, tracking, visualization and interaction, and datasets.

For the first topic, **geometry and appearance digitization**, it was investigated how scenes with heterogeneous appearance can be acquired. A pipeline was realized, which enables a multimodal acquisition and reassembling technique. With this method, it is possible to quantitatively compare photorealism in renderings with actual photos, which also allows to estimate appearance parameters using analysis by synthesis.

Prior work has addressed the challenges of generating and evaluating photorealistic results of physical scenes, but often renderings are not quantitatively compared. With the contributions from this thesis, we have revitalized the quantitative assessment of photorealism, and to the best of our knowledge, this is the first work to do so of heterogeneous scenes requiring multimodal acquisition techniques.

The contribution also provides an insight into a series of technical and practical challenges concerning geometry and appearance acquisition of heterogeneous scenes including glass specifically. While the resulting renderings are good, but not perfect, the contribution allows to measure the errors in the resulting renderings. This allows for further improvements that can then be re-rendered. Future research in this direction should go towards improving how we empirically evaluate geometric and appearance acquisition techniques.

Another contribution to the first topic is an investigation of the ability to capture appearance using a structured light scanner. The method uses simulations and perceptual measures to evaluate the results. Again, the importance of synthesis as a tool for evaluation is highlighted with this contribution. Future work in the direction of this research is obviously to build the proposed system and compare the simulated results, to what can be achieved with a physical implementation of the system.

Finally, the idea of synthesized data was applied as training data to a CNN. It is ideal if a complex model can be applied to generate synthetic images that can be used as a replacement for real photos. It can be very time consuming to acquire large image datasets, and computer rendering can help to speed up this process. Future work in this direction is to further investigate the potential of using CNNs for geometry estimation.

The contributions from the first topic prove that uniting techniques from the areas of Computer Vision and Computer Graphics is necessary to obtain full insights when going from a physical to a digital model.

For the second topic, **tracking**, an innovative technology-to-society solution was realized based on the SfM algorithm, enabling gaze mapping from mobile eye trackers of an AOI. The specific case is defined in Case 2, and the solution provides sufficient accuracy. The method has also contributed to the foundation for an industrial application, and thereby it proves that this work is applicable. Additionally, it provides an extra layer of information, as relative position estimation to an AOI is possible with this method. This information has previously not been directly available from the gaze trackers and provides valuable information to gaze analysis. Future work in the direction of this research, is to improve the tracking algorithm so that it only has to rely on the data provided by the gaze trackers, i.e., without reference images.

For the third topic, **Visualization and Interaction**, a method for visualizing 3D attention heat maps was realized and relates to the work described just above. The contribution provides a technique to improve the way gaze data from mobile gaze trackers is visualized, and while other similar methods exist, it is important to emphasize the defining constraints and the baseline for comparison

of the specific case (Case 2) - both with respect to the contribution to gaze tracking and visualization. Future work in the direction of this research is to further investigate and prove the benefits of this visualization method.

Furthermore, VR was explored as a method for interacting with realistic appearance and geometry data. Two projects was carried out, which can import real-time 3D point cloud data into VR and visualize measured appearance (BRDF) and geometry respectively. Both projects contribute to a unique way of visualizing and interacting with complex data and hopefully motivate the idea of utilizing interactive technologies as a tool for similar tasks. Future work in the direction of this research, is to include more types of physical visualizations. One can imagine to interact with real-time complex meshes or voxels, or to paint with temporally changing BRDFs. Furthermore, existing and future HCI technologies should consistently be evaluated.

For the fourth topic, **datasets**, a series of data has been released regarding the above-mentioned research projects. Much work is required to design and acquire datasets in general, which is often forgotten. Accordingly, the published data is considered a significant contribution. As a matter of fact, for all of the above-mentioned research projects, it has to some extent been necessary to acquire data. While not all of this data is published, it is still a remarkable amount of work put into the data collection, and the insights provided by the different research projects stem from the work that has been put into collecting the data. The data provided of reference geometry of physical glass is considered an exceptionally valuable contribution. Such data is rarely found and can be used by others to further improve on the presented results, or for other research projects including refractive geometry.

In short, the following list of highlighted contributions summarize the work of this thesis:

- A pipeline for capturing multimodal scenes and produce realistic renderings comparable on a pixel to pixel basis, using off-the-shelf techniques. Providing insights into what challenges are associated with such a digitization process, and the complexity of achieving realistic appearance. (Contribution A)
- A method to evaluate and improve upon appearance parameters in a multimodal scene using analysis by synthesis, which allows for quantitative assessment of appearance and geometry. Providing insights in which parameters are dominant when aiming for a one-to-one photorealistic image. (Contribution A)
- A comparative study of design parameters for reconstruction of BRDFs in

a structured light scanner. (Contribution B)

- A technology to society implementation of AOI tracking and automated gaze mapping for mobile eye tracking, also enabling 3D attention heat maps. (Contribution C)
- A VR application allowing for visualization and interaction with point cloud sequences. (Contribution D)
- A VR application for visualization of BRDFs on 3D models, enabling inspection of BRDFs under varying light conditions, and an intuitive and lifelike way to paint 3D models in the digital domain. (Contribution E)
- A dataset with images and reference geometric data of scenes with glass objects.(Contributions F and A)
- A large synthetic dataset of refractive objects for learning algorithms. *Currently not publicly available.* (Contribution H)

The list of contributions proofs that a set of vision based challenges concerning HCI has been addressed and thoughtfully analyzed. Even though not all of the research topics has been investigated in equal depth, they still contribute with state of the art methods and prepares the grounds for future research and investigation.

On an abstract level, this thesis contributes to providing insights on going from the physical to the digital domain using recent Computer Vision and Computer Graphics methods. Additionally, this thesis also investigates interaction and visualization techniques that can ease the way humans can interact with the digital domain. Consequently, it can be concluded that this thesis contributes with improved visual human-computer interaction towards a smooth and flawless experience.

Contribution A

Scene reassembly after multimodal digitization and pipeline evaluation using photorealistic rendering

Scene reassembly after multimodal digitization and pipeline evaluation using photorealistic rendering

JONATHAN DYSSEL STETS[1,†], ALESSANDRO DAL CORSO[1,†], JANNIK BOLL NIELSEN[1], RASMUS AHRENKIEL LYNGBY[1], SEBASTIAN HOPPE NESGAARD JENSEN[1], JAKOB WILM[1], MADS BRIX DOEST[1], CARSTEN GUNDLACH[2], EYTHOR RUNAR EIRIKSSON[1], KNUT CONRADSEN[1], ANDERS BJORHOLM DAHL[1], JAKOB ANDREAS BÆRENTZEN[1], JEPPE REVALL FRISVAD[1,*], AND HENRIK AANÆS[1]

[1] *Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, 2800 Kongens Lyngby, Denmark*
[2] *Department of Physics, Technical University of Denmark, Fysikvej, 2800 Kongens Lyngby, Denmark*
[†] *Joint primary authors*
[*] *Corresponding author: jerf@dtu.dk*

Transparent objects require acquisition modalities that are very different from the ones used for objects with more diffuse reflectance properties. Digitizing a scene where objects must be acquired with different modalities, requires scene reassembly after reconstruction of the object surfaces. This reassembly of a scene that was picked apart for scanning seems unexplored. We contribute with a multimodal digitization pipeline for scenes that require this step of reassembly. Our pipeline includes measurement of bidirectional reflectance distribution functions and high dynamic range imaging of the lighting environment. This enables pixelwise comparison of photographs of the real scene with renderings of the digital version of the scene. Such quantitative evaluation is useful for verifying acquired material appearance and reconstructed surface geometry, which is an important aspect of digital content creation. It is also useful for identifying and improving issues in the different steps of the pipeline. In this work, we use it to improve reconstruction, apply analysis by synthesis to estimate optical properties, and to develop our method for scene reassembly.

OCIS codes: *(150.4232) Multisensor methods; (150.6910) Three-dimensional sensing; (150.1488) Calibration; (160.4760) Optical properties; (290.1483) BSDF, BRDF, and BTDF; (330.1690) Color.*

This is the authors' version of the work. The definitive version is available at https://doi.org/10.1364/AO.56.007679

1. INTRODUCTION

Several research communities work on techniques for optical acquisition of physical objects and their appearance parameters [1–5]. Thus, we are now able to acquire nearly any type of object and perform a computer graphics rendering of nearly any type of scene. The range of applications is broad and includes movie production [2], cultural heritage preservation [3], 3D printing [4], and industrial inspection [5]. A gap left by these multiple endeavors is a coherent scheme for acquiring a scene consisting of several objects that have very different appearance parameters, together with the reassembly of a digital replica of such a scene. Our objective is to fill this gap for the combination of transparent and opaque objects, as many real world scenarios exhibit this combination. An example is a living room, like the one rendered in Fig. 1 (right). We propose a pipeline for acquiring and reassembling digital scenes from this type of heterogeneous real-world scenes. In addition, our pipeline closes the loop by rendering calibrated images of the digital scene that are commensurable with photographs of the original physical scene (see Fig. 1, left). This allows for validation and fine-tuning of appearance parameters. The quantitative evaluation we get from pixelwise comparison of rendered images with photographs is a great improvement with respect to validation of the acquired digital representation of the physical objects.

When addressing the problem of acquiring a heterogeneous scene, there is an infinite variety of scenes and object types to choose from. So, to make our task feasible, we focus on scenes that combine glassware and non-transparent materials, more specifically, white tablecloth and cardboard with a checkerboard pattern. We made these choices as glass requires a different acquisition modality, the tablecloth bidirectional reflectance distribution function (BRDF) is spatially uniform but not necessarily simple, and the cardboard has simple two-color varia-

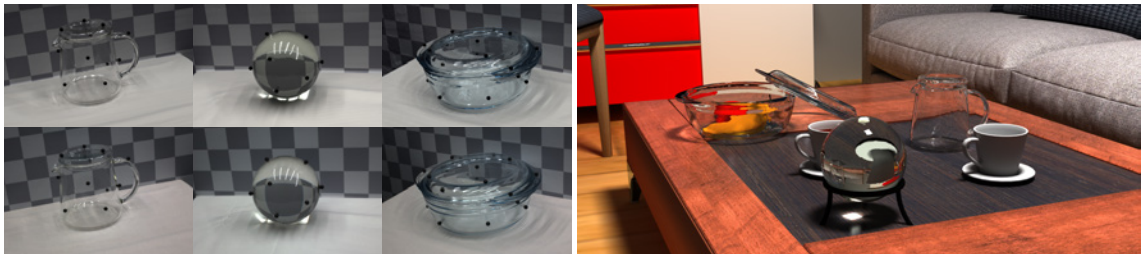


Fig. 1. To the left, we compare rendered images (top) with photographs (bottom). More views are available in Appendix A. The scenes to the left were digitized using our pipeline and include both glass objects and non-transparent objects (tablecloth and backdrop). To the right, we exemplify the use of our pipeline for virtual product placement using our digitized glass objects, with estimated optical properties and artifact-reduced removal of markers.

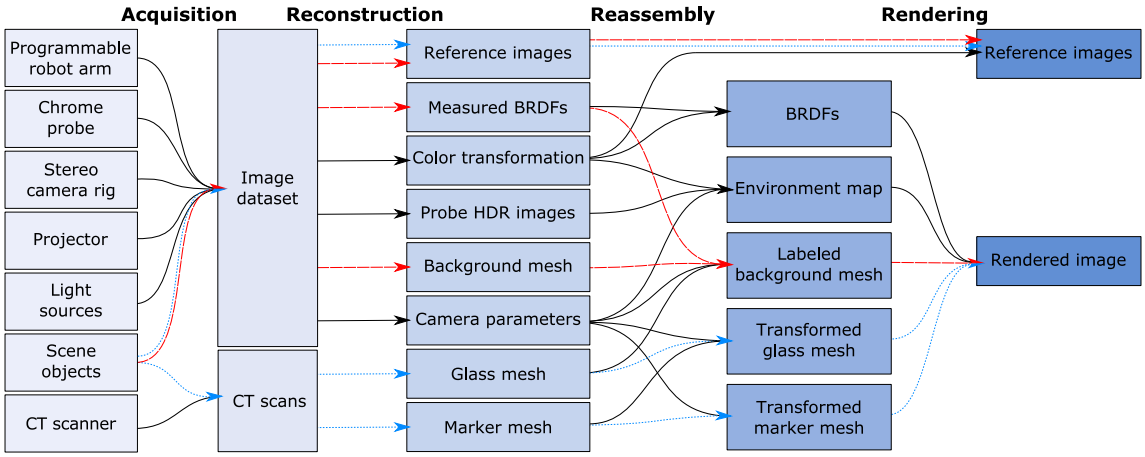


Fig. 2. Overview of our digitization pipeline in four main stages: acquisition, reconstruction, reassembly, and rendering. A video presentation of our pipeline is available in supplementary Visualization 1. Colored arrows show the path through the pipeline of transparent objects (dotted blue) and non-transparent objects (dashed red).

tion. The latter is particularly useful for observing how light refracts through the glass. The chosen case is also of particular interest, since glass is present in many intended applications of optical 3D acquisition. Considering the highly multidisciplinary nature of our work, we have released our dataset (http://eco3d.compute.dtu.dk/pages/transparency). This facilitates further investigation by other researchers of the different steps of our pipeline with the possibility of a quantitative feedback at the end of the process.

A. Related Work and Contributions

Researchers occasionally compare renderings with photographs to provide a qualitative verification of a presented rendering technique. The work by Phong [6], Goral et al. [7], and Takagi et al. [8] are early examples of this trend. A procedure to bring a rendered image close to a photograph was first presented by Meyer et al. [9]. In this work, likeness of images was evaluated perceptually by human observers. Pixelwise comparison of photographs with rendered images is surprisingly uncommon. The few examples we have found are by Rushmeier et al. [10], Karner and Prantl [11], Pattanaik et al. [12], and Jones and Reinhart [13, 14]. These examples build on the rendering framework described by Greenberg et al. [15]. Employing such a framework for more complex scenes is a long and tedious process [16]. The key issue is that a scene specification is expected as an input.

Several problems arise as a result of not having correspondence between the physical and the digital scene. Misalignment due to inaccurate scene and viewing geometry and inaccurate orientation of the lighting environment are some of the essential problems identified in previous work [17, 18]. One way to deal with this problem is to calculate error for image patches when evaluating results [13, 19, 20]. As opposed to this, our digitization pipeline (Fig. 2) provides both reference photographs and correspondingly calibrated scene and viewing geometry so that pixelwise comparison becomes meaningful.

Pixelwise comparison of rendered images with photographs is not only useful for quantifying the photorealism of a rendering in terms of error measurements. We find it particularly useful for improving the digitization pipeline. The fact that our pipeline enables quantitative evaluation led us to more specific contributions in its different steps. These contributions are mostly in the reassembly and are as follows. (a) A cross-modality marker-based placement approach, enabling accurate placement of objects scanned with one modality into scenes scanned with

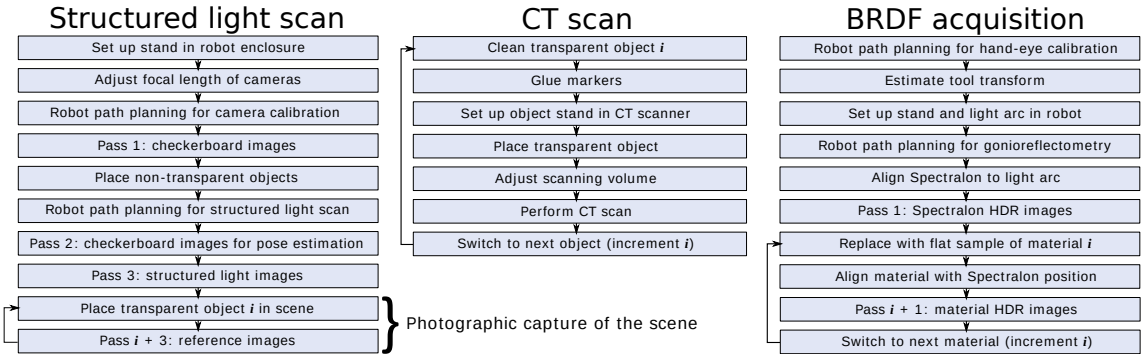


Fig. 3. Our workflow for scanning the geometry of non-transparent objects and collecting reference images (left), for scanning the geometry of transparent objects (middle), and for measuring material reflectance properties (right).

another modality. (b) A soft object deformation technique dealing with surface intersections after object placement, which is critical for scenes containing transparent or translucent objects. (c) A micropolygon labeling approach for assigning BRDFs to acquired geometry. (d) A color calibration scheme enabling use of spectral optical properties for calculating reflectance, transmittance, and absorption. (e) Perspective unwrapping of mirror probe images to improve precision when the environment is not very distant. (f) Use of analysis by synthesis for fine-tuning physics-based optical properties.

Digitization is most often unimodal and tailored toward objects with a specific type of surface reflectance behavior [1]. While unimodal techniques are becoming more versatile [21–23], objects with a transparent material like glass still pose challenging problems. Their reflectance behavior is so different that they require an entirely different modality, such as computed tomography (CT) [24]. The transparent object must then be removed from the scene to be scanned elsewhere. In the meantime, the surrounding scene can be scanned with a more common technique. However, as the transparent object takes most of its appearance from its surroundings, it must be repositioned in the surrounding scene (physically and digitally) if we are to take reference images for comparison with rendered images. The purpose of our scene reassembly is to address this type of issue.

Our digitization technique is multimodal. Currently, such techniques seem to exist only in the context of sensor fusion [25–27]. Here, the goal is to optimize reconstruction by fusing data from different sensor modalities with complementary characteristics. Even so, the different modalities see the same object and thus work for materials with a similar reflectance behavior. The challenge is then mostly in registration of the scans. In their final remarks and suggestions for future work, Weinmann and Klein [1] discuss possible ways of combining multiple techniques tailored to different types of surface reflectance. Our pipeline is a different way to take a step in this direction.

In summary, our work makes it possible to perform multimodal digitization and scene reassembly in such a way that rendered images of the reassembled scene can be quantitatively compared to photographs of the original. This enables us to provide the first empirically founded investigation of the appearance accuracy of objects digitized using a non-optical scanner.

2. DIGITIZATION PIPELINE

We divide our pipeline into four stages: (1) acquisition, (2) reconstruction, (3) reassembly, and (4) rendering. Figure 2 provides an overview. As illustrated, transparent objects (dotted blue arrows) and non-transparent objects (dashed red arrows) take different paths through the pipeline. The acquisition stage includes structured light scanning of non-transparent objects, CT scanning of transparent objects, gonioreflectometric reflectance measurements, and photographic capture of environment, color chart, and scene reference images. Figure 3 provides details of our workflow in these acquisition steps (except the simpler captures of environment and color chart). The second stage includes reconstruction of surface meshes, material BRDFs, and color space. The third stage is reassembly of the digital scene consisting of geometric objects, material appearance properties, and environment map. The fourth and final stage is rendering and comparison with reference images.

Our acquisition stage requires an elaborate hardware setup. We assemble the physical scene in a black light-proof enclosure. This has five LED light tubes for scene lighting, which we capture by high dynamic range (HDR) imaging of a light probe. To acquire non-transparent geometry inside this enclosure, we use a structured light scanner consisting of a toe-in stereo camera rig and a light projector mounted on a robotic arm [28, 29]. We chose a converging camera configuration (toe-in) to increase the overlap of the fields of view so that we get a denser point cloud per stereo view. Together with an LED based illumination arc, we also use this camera rig with exact control for measuring isotropic BRDFs. For transparent objects, we use a CT scanner. In the following subsections, we describe the individual steps of the pipeline with focus on details required for reproducibility and on non-standard techniques that we introduce.

A. Camera Calibration and Settings

The camera system is calibrated using a standard technique [30]. Our calibration board is an 11 by 12 black-and-white checkerboard. For the intrinsic calibration (Pass 1 of Fig. 3, left), we include a large variety of views to estimate good lens distortion coefficients. To facilitate stereo calibration, we also ensure that both cameras have the calibration board fully in view. For extrinsic calibration (Pass 2 of Fig. 3, left), we balance good coverage of the scene and good coverage of the calibration board. Since we cannot change the camera system while collecting data, we

chose a small aperture to ensure that background and projected structured light patterns are always in focus from all views. The full setup is in a dark room environment to eliminate external light, so we use a long shutter time (600 ms) to obtain sufficient exposure. A slight noise component is present in the images, but this is considered negligible. Finally, we use the estimated distortion coefficients to remove distortion from all images in the dataset so that subsequent algorithms may assume a pinhole camera model.

To avoid any compression or manipulation of the images by the camera software, in particular automatic color correction, we read the raw sensor data directly. We use bilinear interpolation to reconstruct RGB images from the raw Bayer pattern images. By doing this, we obtain a consistent RGB color space. Moreover, the raw sensor data is linear and correlates directly with radiometric quantities, which allows for better BRDF and environment map estimation in later stages of our pipeline.

We capture radiometrically relevant parts of our dataset in HDR by stacking multiple exposures [31]. More specifically, we stack 11 exposures at one-stop intervals ranging from 1 to 2048 ms. For the other parts of the dataset, we capture a single image at an exposure time of 600 ms.

B. Surface Reconstruction from Structured Light

We use a standard Gray code structured light approach to generate raw point clouds for a scene [32, 33]. With camera parameters from the calibration, we transform these point clouds into the same world coordinate system.

To reconstruct one connected triangle mesh from the point clouds, we merge them into a single point cloud and perform screened Poisson reconstruction with trimming and an octree depth of nine [34]. This technique requires point normals, so before the merging we generate normals for each point cloud as follows. We resample the point cloud down to 100,000 vertices via Poisson disk sampling [35] and then compute normals via planar fitting to a nearest neighborhood of 500 points ($\sim$16 mm radius). We then reorient all the normals according to the location of one of the cameras and transfer them back onto the original point cloud. This procedure ensures smooth continuous normals, necessary for a good performance of the mesh reconstruction algorithm. As we rely on smoothing, we cannot reconstruct features in the mesh with the same physical size as the alignment error accumulated from structured light and calibration. The aim of the chosen constants was to preserve features by striking a balance between too noisy and too smooth. The operability of the pipeline is however not sensitive to the choice of these constants.

C. Material BRDF Reconstruction

We assume that all non-transparent materials in the scene are opaque and isotropic, so we model their reflectance properties by BRDFs. To acquire a BRDF, we combine traditional canonical gonioreflectometric sampling [36] with a BRDF interpolation (reconstruction) technique [37]. We follow the workflow outlined in Fig. 3 (right). A light arc illuminates material samples from 11 unique inclinations, evenly distributed from 7.5° up to 90° with 7.5° steps. We place a flat material sample at the center of the circle partly traced by the light arc. Using the cameras mounted on the robot, we then measure radiance reflected by the sample across one octant of a sphere. The center of this sphere coincides with that of the light arc, while its radius is slightly larger to avoid collision between the robot and the arc. The robot moves in steps of 7.5° and captures 11 HDR images of the sample per step, one for each light direction. In total, this yields 2,783 HDR images per material. We avoid tangential and zenith viewing directions (90° and 0°, respectively). In the former case, no reflected radiance should be visible, while in the latter the light arc occludes the view of the sample.

The 2,783 observations are too few to faithfully represent the BRDF of a material in a photorealistic rendering. We need an interpolation scheme to fill the entire $(90 \times 90 \times 180)$ Mitsubishi Electric Research Laboratories (MERL) format BRDF look-up table [38]. The reconstruction method by Nielsen et al. [37] is our interpolation scheme. First, we use each of the 100 BRDFs in the MERL-dataset [38] as sample points in a $90 \cdot 90 \cdot 180 = 1{,}458{,}000$ dimensional space. The nonlinear mapping of Nielsen et al. [37] is then applied to each of the samples. The mapped samples are ordered as rows of a matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ where m is the number of BRDF samples and d is the dimension of the space. The zero-mean matrix is computed as $\mathbf{X} - \bar{\mathbf{x}}$, with $\bar{\mathbf{x}}$ being the sample mean. From this, the singular value decomposition $\mathbf{X} - \bar{\mathbf{x}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is used to compute the eigenvectors and eigenvalues of the covariance matrix of $\mathbf{X} - \bar{\mathbf{x}}$, which are given as the columns of $\mathbf{V}$ and the diagonal elements of $\mathbf{\Sigma}$, respectively. This is effectively a principal component analysis (PCA), where the eigenvectors are the principal components. A matrix composed of the scaled principal components as columns are computed as $\mathbf{Q} = \mathbf{V}\mathbf{\Sigma}$.

Now, the full BRDF can be reconstructed from this principal component space by projection. Let $\mathbf{x}' \in \mathbb{R}^n$ be n BRDF observations measured for a given material. Then, let $\bar{\mathbf{x}}' \in \mathbb{R}^n$ be the mean values and $\mathbf{Q}' \in \mathbb{R}^{n \times k}$ be the scaled eigenvectors corresponding to the direction pairs of those n observations. A vector $\mathbf{c}$ which spans the full space can be constructed by finding the linear combinations of principal components that best approximate the n observations. We do this by solving the linear least-squares optimization problem given by

$$\begin{aligned} \mathbf{c} &= \arg\min_{\mathbf{c}} \|(\mathbf{x}' - \bar{\mathbf{x}})' - \mathbf{Q}'\mathbf{c}\|^2 + \eta\|\mathbf{c}\|^2 \\ &= (\mathbf{Q}'^T\mathbf{Q}' + \eta\mathbf{I})^{-1}\mathbf{Q}'^T(\mathbf{x}' - \bar{\mathbf{x}}'). \end{aligned}$$

Note that by adding a penalty η to the norm of $\mathbf{c}$, this effectively becomes a Tikhonov regularized least squares. Now, the full, mapped BRDF is reconstructed as $\mathbf{x} = \mathbf{Q}\mathbf{c} + \bar{\mathbf{x}}$. The inverse of the nonlinear mapping applied to $\mathbf{X}$ is applied to $\mathbf{x}$ to get the actual, unmapped BRDF of the material. The described approach is applied to every single non-transparent material in the scene in order to obtain models of their reflectance properties.

This approach assumes that the MERL database encompasses the class of materials present in the scene. Effectively, this is a practical compromise between dense, unbiased, canonical BRDF sampling and fast, inferred BRDF sampling. This enables us to obtain high confidence BRDFs in a matter of a few hours.

D. Surface Reconstruction from CT

In our dataset, we have three glass objects: a sphere, a teapot (pot and lid) and a bowl (bowl and lid), for a total of five pieces. All objects have spherical plastic markers glued onto their outer surface. We CT scan each glass piece to obtain X-ray radiographs and use the CT PRO 3D reconstruction software from Nikon Metrology to obtain a volumetric image for each piece. The resolution of the reconstructed volume is up to 1000^3 voxels. Due to beam hardening, high density differences between materials lead to streak artifacts [39], especially around our markers and at the top and bottom of the objects (see Fig. 4). We account for these artifacts in the volumetric segmentation.

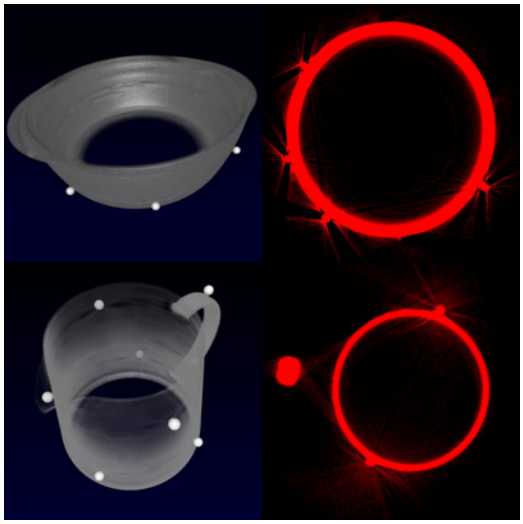


Fig. 4. CT scans of the bowl (top row) and the teapot (bottom row) with markers glued onto them. In the left column, visualized using a 1D transfer function. Note the different density of the markers. In the right column, a slice scaled to display streak artifacts.

From a CT scan, we generate two triangular meshes with vertex normals: one for the glass object and one the plastic markers. Figure 5 provides an overview of our procedure. We start with the markers, which appear as elements of higher density in the scan. We preprocess the scan by clamping all the values under a certain threshold to zero and then create a mesh using dual contouring [40]. Generating the glass mesh is more cumbersome. We also use dual contouring in this case, but because of the streak artifacts (Fig. 4) it is not possible to isolate the glass mesh via a threshold. Instead, we use a lower threshold that only removes noise, then estimate the marker positions, and use these to remove the markers from the glass mesh.

To estimate marker positions, we determine a series of center/radius pairs $(\mathbf{c}_i, r_i)$ by fitting a multi-sphere model to the marker mesh vertices using a tuned random sample consensus (RANSAC) algorithm [41]. We then carve a hole by excluding all the triangles that are inside a sphere with center $\mathbf{c}_i$ and radius $(1+\epsilon)r_i$, where ϵ is usually in the 0.5 to 0.75 range. We store the marker positions $\mathbf{c}_i$ so that we can use them to transform from the local coordinate system of the glass object to the world coordinate system (see Section F).

After removing the markers, the glass meshes still have aliasing artifacts. To deal with this issue, we first decimate the mesh down to 1% of the original vertices via quadric edge collapse. The holes are then easy to close by identifying the edge loops surrounding each hole and filling these with triangles. We then introduce a subdivision-decimation loop with alternating $\sqrt{3}$-subdivision [42] and decimation to 33% of the original vertices. We perform this subdivision-decimation operation four times to obtain a cleaned mesh. The decimation removes unwanted high frequency features from the mesh. Thus, we generate smooth meshes at the cost of some geometric precision. We are again trying to strike a balance between reconstruction error and too much smoothing. In Section 4, we compare our method with a different cleaning procedure that better preserves geometry.

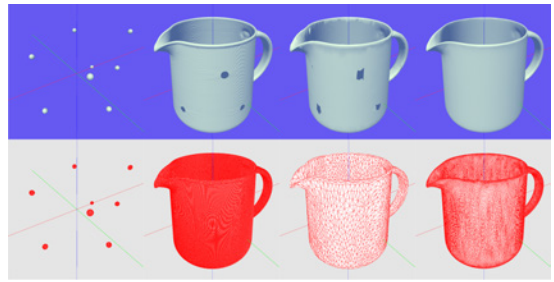


Fig. 5. Reconstruction from CT with stages illustrated using Phong shading (top row) and wireframe shading (bottom row). After estimating the marker mesh (first column) and fitting spheres to the markers, we reconstruct the object mesh (second column). To eliminate noise, we first simplify the mesh (third column) and then close the holes and apply our subdivision-decimation loop to get the final object mesh (fourth column).

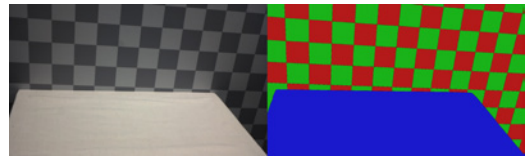


Fig. 6. Labeling of the image to the left results in the label image to the right. Each color in the label image represents a label that we assign a BRDF to. The black edges between labels indicate areas where we apply a nearest neighbor method.

E. Scene Reassembly for Non-Transparent Objects

Two operations are necessary to prepare the background mesh for rendering: labeling and deformation. In the labeling, our objective is to identify BRDFs and label each face of the mesh with a BRDF. Assuming a scene with a small number of known BRDFs, we apply edge detection and watershed on the images of the scene to segment BRDF boundaries. Shadows, specular highlights, and different viewing angles of the scene complicate fully automatic BRDF identification. Our approach gets us most of the way, but we manually correct any residual misclassification. Figure 6 shows a label image produced by our labeling technique.

The label images can be used in multi-view projective texturing of the background mesh. However, we would like to precompute the view and label selection instead of doing it millions and millions of times while rendering. To avoid uv-unwrapping of the mesh for storing precomputed labels, we take an approach inspired by micropolygon rendering [43]. We project each vertex of a face onto the label images of the scene and select the face BRDF according to the image label that most of the face vertices were projected to. If a vertex projects to an unknown label, we resolve it by a nearest neighbor search. Since faces around material boundaries overlap multiple materials,

No subdivision One subdivision Two subdivisions

Fig. 7. Subdividing the mesh dissolves unwanted boundary sawtooth artifacts that originate from the BRDF labeling.

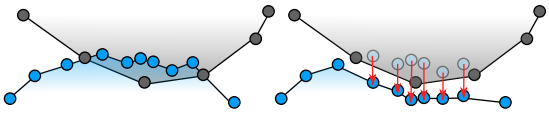


Fig. 8. Deformation of background mesh, where we push the background vertices down to avoid mesh intersection.

we get sawtooth artifacts. We dissolve these by subdividing the mesh until the rendered triangles are smaller than the surface area observed in a pixel, see Fig. 7.

When applying physically based rendering, we observed intersections between background scene and glass meshes. This could be due to small errors in reconstruction and positioning, or perhaps the harder glass objects press down the tablecloth when placed for reference imaging. It causes significant visual artifacts since the rendering exposes all surfaces of a transparent object. To eliminate these artifacts, we accommodate the hard object (glass) by deforming the soft object (tablecloth), see Fig. 8. To deform the soft object, we need a "down" direction in which to push the vertices. We first find contact vertices. These are vertices in each mesh that are close to any vertex of the other mesh. We consider vertices close if the distance between them is less than 7% of the bounding box diagonal of the hard object. Using least squares regression, we fit a contact plane to the contact vertices of the soft object. We set the sign of the contact plane normal so that the upper half-space contains the center of the hard object bounding box. Projection of a contact vertex to the normal of the contact plane then measures the height of the vertex. For each soft object contact vertex $\mathbf{x}$, we find the nearest hard object contact vertices and push $\mathbf{x}$ down below the lowest one of these.

F. Scene Reassembly for Transparent Objects

To reposition the glass objects in the scene, we rigidly transform the meshes reconstructed from CT to the world coordinate system of the background mesh. We obtain this transformation by matching markers in the stereo images with the marker coordinates $\mathbf{c}_i$ computed during reconstruction from CT (see Section D).

To find the markers, we employ a size invariant circle Hough transform [44]. This works well for our dataset, where the markers show high contrast against their surroundings. We match markers in the left and the right images via Sampson distance [45]. Using this technique, markers on the same epipolar line lead to false positives, so we manually inspect the result. We also manually discard detected markers that are visible through the glass, as the refraction would lead to incorrect positioning. Markers in both stereo images with no match are discarded. The result is a set of matched markers in image coordinates as seen in Fig. 9 (bottom left). We then triangulate the matched markers

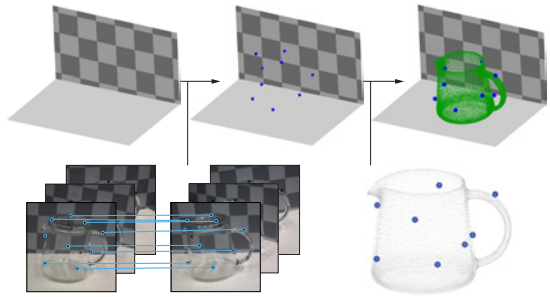


Fig. 9. Repositioning a CT scanned object in the background scene. We identify and match the markers in the stereo image pairs and calculate their corresponding 3D points. Pairing these with marker coordinates from the CT scans, we transform the CT scanned piece of an object into the world coordinate system.

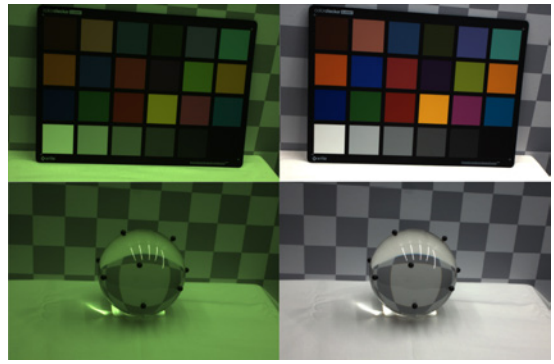


Fig. 10. Color calibration: raw images (left) and color corrected images (right). The camera sensor is particularly sensitive to green.

from the stereo views and gather them in clusters of 3D points. We remove outliers via their distance from the cluster centers, and for each cluster we select the point with the lowest reprojection error. An example of the points and clustering is shown in Fig. 9 (top middle).

We manually pair the 3D marker coordinates from the images with the marker coordinates $\mathbf{c}_i$ from the CT scans. We perform Procrustes analysis [46] on the two point sets, excluding reflection, since we assume a rigid transformation applied to each vertex of the mesh. The bowl and the teapot are composed of multiple pieces. For these objects, we compute the transformation individually for each piece. The result of the object transformed into the scene is shown in Fig. 9 (top right). We found that in order to have low error in the transformation the chosen markers should sample the surface evenly and be visible from most views.

G. Color Calibration

Images are only quantitatively comparable if they live in the same color space. Thus, we must ensure that our radiometry-dependent data, namely reference images, environment map, and BRDFs, are in the same color space. We do this by imaging

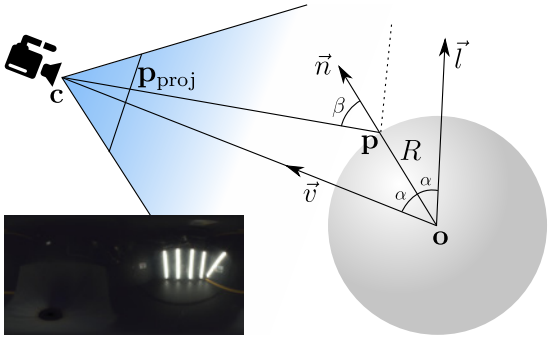


Fig. 11. Unwrapping of a spherical probe. We know the sphere radius R from specification, the camera position $\mathbf{c}$ through calibration, and the sphere center $\mathbf{o}$ by triangulation. Radiance at $\mathbf{p}_{\text{proj}}$ in our image then corresponds to the environment map direction $\vec{l}$. The result for the robot enclosure is in the lower left corner in latitude-longitude panoramic format (here tone-mapped).

a color chart of precisely known colors. More specifically, we use second degree root-polynomial color correction [47] based on a 24 patch ColorChecker Classic from X-Rite. This provides a matrix that transforms from camera RGB to XYZ, where we assume illuminant D50 when specifying the XYZ values of the colorchecker. With the assumption of illuminant D50, we can transform colors to the CIE L*a*b* color space and then compute color difference using the ΔE_{00} metric [48]. We use this to refine our result by minimizing ΔE_{00} using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [49]. The result is in Fig. 10. The average color difference is $\Delta E_{00} = 1.97 \pm 1.21$, which is larger than 1 JND (just noticeable difference) [50], but we find it acceptable.

Since we work with glass objects (and chrome, see Section H), we need refractive indices to determine reflectance, transmittance, and absorption properties. Refractive indices can be found per wavelength in tables of research papers. To use such spectral optical properties together with our trichromatic image data, we integrate them to CIE RGB using the CIE RGB color matching functions listed by Stockman and Sharpe [51]. It is important to normalize these functions [52] and to use RGB rather than XYZ [53]. This is because a refractive index is not a color, but rather a quantity that in trichromatic representation should resemble a sparse sampling of the spectrum. Thus, as recommended by other authors [54], we choose CIE RGB as our rendering color space. After transforming our image data from camera RGB to XYZ, we therefore convert them to CIE RGB [55]. As a final step, we apply Bradford chromatic adaptation [50], adapting to the originally assumed illuminant D50, so that renderings and reference images get closer to real life appearance.

H. Environment Lighting

To capture the lighting observed in the reference images, we use a method similar to the mirror probe technique [56]. However, we use a pinhole camera model for probe image unwrapping instead of the standard orthographic model. Our pipeline enables this as we have a calibrated camera and know its position relative to the photographed mirror probe. With the pinhole model, we obtain a more precise estimate of the environment lighting. The environment map is generated from HDR images and stored in latitude-longitude panoramic format [50]. We use a polished grade G100 chrome bearing ball as mirror probe.

An environment map represents an infinite area light and maps a direction to a texture element (a texel). To do unwrapping, we map each texel direction $\vec{l}$ to the corresponding pixel position $\mathbf{p}_{\text{proj}}$ in a light probe image. Given the configuration illustrated in Fig. 11, we have

$$\vec{v} = \frac{\mathbf{c}-\mathbf{o}}{\|\mathbf{c}-\mathbf{o}\|},\ \vec{n} = \frac{\vec{v}+\vec{l}}{\|\vec{v}+\vec{l}\|},\ \mathbf{p} = \mathbf{o} + R\vec{n},\ \mathbf{p}_{\text{proj}} = \mathbf{M}\,[\mathbf{p}^T\ \ 1]^T,$$

where camera matrix $\mathbf{M}$ and camera position $\mathbf{c}$ are available from our calibration. The radius of the sphere R is available from the bearing ball specification, and we find the center of the sphere $\mathbf{o}$ by manually annotating the sphere and then triangulating it. We assume that the distance to the actual light along $\vec{l}$ is equal to the distance between camera and sphere $\|\mathbf{c}-\mathbf{o}\|$. This assumption works well in practice, leading to an error smaller than the uncertainty of $\mathbf{o}$ caused by the triangulation. With the original orthographic camera model, we can reconstruct the lighting for all directions except one ($-\vec{v}$). In our model, we cannot reconstruct the lighting for a set of directions ($\vec{n}\cdot\vec{v} \le R/\|\mathbf{c}-\mathbf{o}\|$), so we set them to black. Since we do our unwrapping in world space, we can combine contributions from multiple camera views with no need to align them afterwards.

The environment map is color corrected according to Section G, which enables us to correct for the angularly dependent reflectance of chrome. The correction is to divide by Fresnel reflectance, which we compute during unwrapping. As input for Fresnel's equations, we use the angle β between $\mathbf{c}-\mathbf{p}$ and $\vec{n}$ and the complex refractive index of chrome [57] converted from spectrum to CIE RGB. The result is shown in the inset of Fig. 11.

I. Rendering

We render images using progressive unidirectional path tracing [58, 59] implemented in OptiX [60]. The captured HDR environment map is the sole light source in our scene [56]. When rendering non-specular materials, we importance sample the environment map to get direct illumination and use sampling of a cosine-weighted hemisphere to get indirect illumination. From our labeling, we have one BRDF attached to each triangle in our scene. For non-transparent objects, we use our measured BRDFs tabulated in the MERL format [38]. To terminate paths probabilistically, we use Russian roulette based on the bihemispherical reflectance of each measured BRDF. This reflectance is calculated in a preprocessing step using Monte Carlo integration. We deal with transparent objects in the usual way, setting reflectance and transmittance according to Fresnel's equations of reflection and Bouguer's law of exponential attenuation. Given their small surface, we were unable to estimate a BRDF for the markers. Instead, we render them as glass with all refracted rays being absorbed.

3. ANALYSIS BY SYNTHESIS

The ability to render images comparable to photographs enables us to use our pipeline for improving parameter estimates through analysis by synthesis. As an example, we need a scaling factor for our HDR environment map as it measures relative radiance [31]. We estimate this factor by taking ratios of references

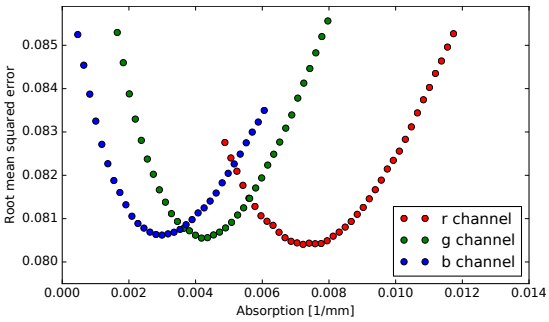


Fig. 12. Analysis by synthesis to estimate absorption of the glass bowl. We run renderings in low resolution and change the absorption in each color channel one at the time. In the case of the bowl, the blue channel is the most sensitive one.

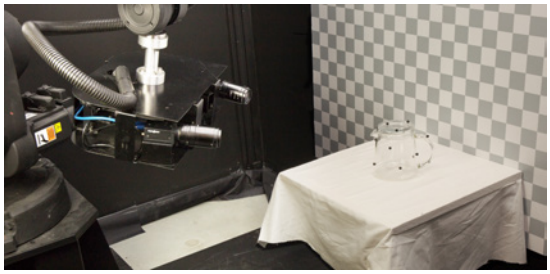


Fig. 13. Scene with checkerboard backdrop, lighting, glass teapot, and stand with table cloth observed by two cameras mounted on a 6-axis industrial robot arm.

and renderings with the background scene alone. Another example is estimating real and imaginary parts of glass refractive indices. As analysis by synthesis is fundamentally ill-posed [61], we take our outset in physics-based initial guesses such as Schott K5 crown glass (sphere and teapot) and soda lime glass (bowl). Spectral refractive indices for these glasses were obtained from an online database (http://refractiveindex.info) and converted to CIE RGB. All parameters were estimated using different views than the ones in our comparisons of renderings with references.

As an example of our analysis by synthesis, we plot the evolution of the root-mean-squared error (RMSE) for different renderings of the glass bowl in Fig. 12. For each rendering, we vary a trichromatic component of the absorption coefficient (which directly relates to the imaginary part of the refractive index). We identify a distinct minimum in the error for each channel, with a slightly larger uncertainty in the red channel. The minimum values in this figure were used in our renderings of the glass bowl. We apply the same analysis to the teapot and the sphere.

Given an initial guess for a parameter, we can employ standard optimization algorithms, defining the RMSE between the reference and the rendering as a cost function to minimize. To reduce rendering times, the evaluation of the cost function can be calculated on a downsampled image or limited to a specific patch of the images. Various general optimization algorithms exist for minimizing expensive cost functions [62].

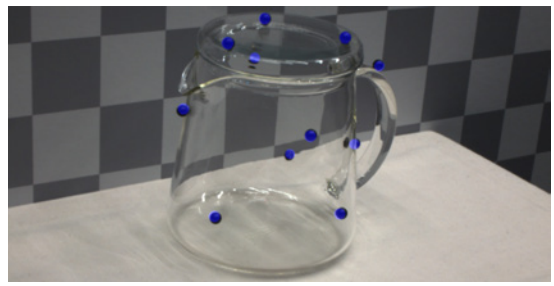


Fig. 14. Markers rendered in blue and added to the reference image to validate marker positions by looking at pixel offsets.

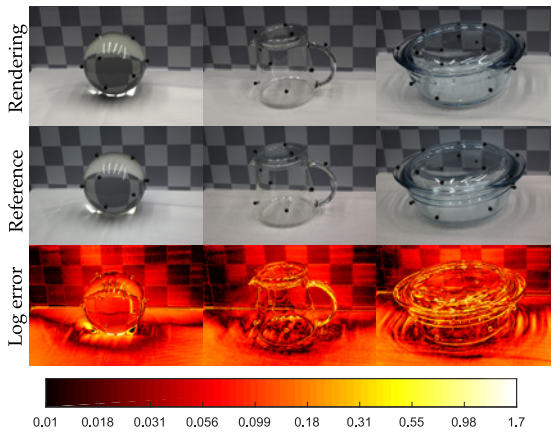


Fig. 15. Pixelwise error for three rendering-reference pairs. Error is the ℓ^2-norm of 32-bit per channel RGB images, visualized using a base 10 logarithmic scale.

4. RESULTS

Our scenes consist of a backdrop, a stand, and a glass object (with markers) placed on the stand. The backdrop is a 30 by 20 white-and-gray checkerboard print on 120 cm by 80 cm semi-matte cardboard and the stand is a tabletop with a white cloth. An example scene is depicted in Fig. 13. We implemented our reconstruction and reassembly procedures as a modular software pipeline and computed all rendered images using our path tracer. As illustrated in Fig. 2 and mentioned in Section G, we color correct both rendered images and reference images to have a meaningful perceptual comparison. Figure 14 compares markers in a reference image with rendered markers to validate our marker positioning. For the teapot, the average distance between the markers from stereo and the transformed markers from CT is 0.43 mm.

Figure 15 presents pixelwise comparisons of reference images and rendered images. The error images allow us to spot subtle differences not easily noticed in a perceptual comparison, such as the slight misalignments in geometry and highlights. As reference photographs were not captured in HDR, we clamp the renderings correspondingly. This means that areas of strong light intensity, such as highlights and intense caustics, appear black in the error images.

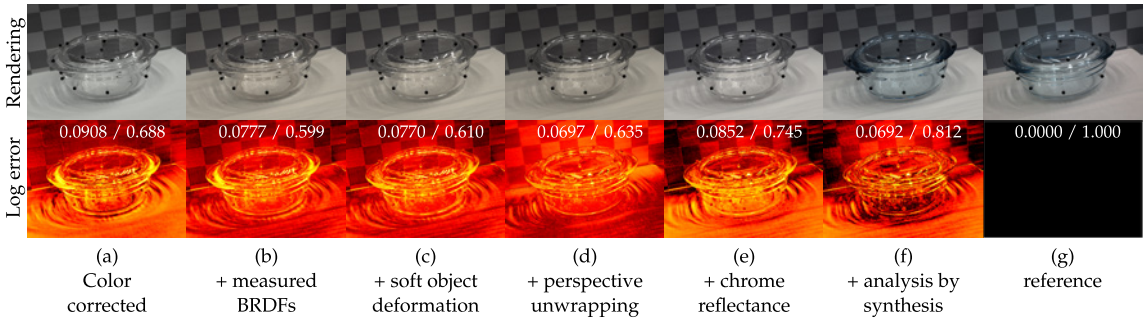


Fig. 16. Qualitative (top) and quantitative (bottom) step-by-step evaluation of our reassembly techniques. The log error images have the same format as in Fig. 15 and the reference photograph is in the rightmost column (g). In each column, we provide root-mean-squared error and structural similarity index (RMSE / SSIM). Both measures attain their best score in our final result (f).

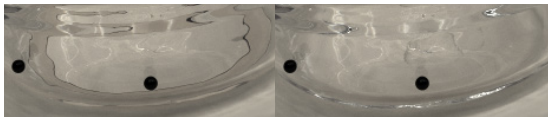


Fig. 17. Zoom-in of Figs. 16 (b) and (c) to emphasize the effect of our background deformation.

Fig. 18. Zoom-in of Fig. 16 (c) and (d) to emphasize the effect of our perspective unwrapping of the environment map.

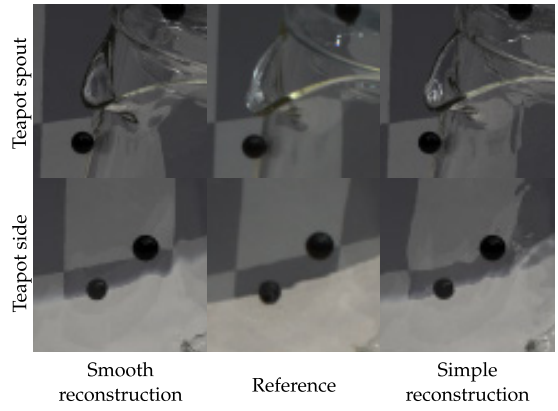
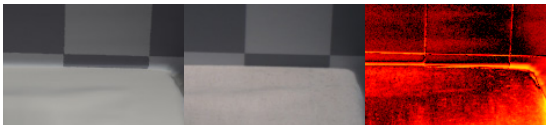


Fig. 19. Trade-off in mesh reconstruction. If we smooth more, we get less distortion in the refractions, but less precision in the mesh geometry. From left to right: Rendering with smoothing, reference image, rendering without smoothing.

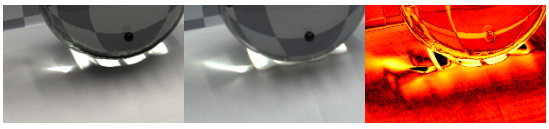
Figure 16 exemplifies the impact on error images of some of our contributions. In Fig. 16 (a), we only reposition the glass object in the background scene and apply color correction (Sections F and G). This means that we use Lambertian materials (with bihemispherical reflectances from the measured BRDFs), an orthographic unwrapping model of the environment map, and no chrome reflectance correction or analysis by synthesis optimization. We compare to the reference image in Fig. 16 (g), with error images as in Fig. 15. Figure 16 (b) shows the impact of using measured BRDFs (Section C), resulting in a more accurate representation of the folds of the cloth in the background scene (top image) and an overall reduction of the error (bottom image). In Fig. 16 (c), we add deformation of the background mesh (Section E), which ensures that the background mesh does not poke through the glass surface (see a close-up in Fig. 17). Additionally, we can see how this improves the error on the lid of the bowl, because of refraction of light in the glass. The next step, Fig. 16 (d), shows the impact of our modified environment map unwrapping (Section H) against the standard orthographic unwrapping rotated according to our camera parameters. A close-up is available in Fig. 18. Our modified unwrapping provides a better shape and alignment of highlights and caustics. Partially due to the assumption of infinitely distant environment light, some alignment artifacts persist. In Fig 16 (e), we show the effect of correcting for chrome reflectance in our environment map reconstruction. Quantitatively, this changes the distribution of the error (bottom image). On the cloth, the exposure increases, exposing the caustics misalignment. On the backdrop, the error reduces. Interestingly, the structural similarity index (SSIM) improves while the RMSE worsens. Finally, in Fig. 16 (f), we use analysis by synthesis to adjust glass absorption. This improves the glass appearance, but it also leads to slight color changes in other parts of the scene due to indirect light paths. Because of this global influence, the analysis by synthesis introduces slightly too much absorption to compensate for the slightly too bright tablecloth.

As an example of how our pipeline can be used to validate existing algorithms, we investigate the case of glass object reconstruction. In Fig. 19, we compare two different reconstruction methods with focus on two parts of the teapot scene. Smooth reconstruction refers to the procedure described in Section D. The other procedure is to simply decimate the reconstructed mesh to 2.5% of the original vertices and apply Taubin smoothing [63]. This removes the high frequencies of the noise but much noise is still present in the midranges leading to wobbly refractions.



Rendering Reference Log error (as in 15)

Fig. 20. Material transitions: error lines along checker edges and along the boundary between tablecloth and backdrop.



Rendering Reference Log error (as in 15)

Fig. 21. Effect of separating markers from glass (refracted light close to marker) and of not accounting for subsurface scattering (dark areas close to caustics).

Our method in Section D reduces far more noise, but this is at the cost of greater changes to the overall shape. We note that a refractive object with a simple geometry is very hard to reconstruct automatically if fidelity and almost no noise are both required.

5. DISCUSSION

Since our pipeline enables us to compare renderings with photographs, we can identify problems in acquisition, reconstruction, and rendering that would otherwise have been hard to find. Camera calibration issues, for example, reveal themselves as error lines along edges (visible in Fig. 20). Color calibration issues reveal themselves as color shift. Such issues led us to more careful camera calibration procedures and the choice of root-polynomial color correction. Qualitative comparisons revealed artifacts in surface reconstruction, mesh intersections calling for deformation, misplacement of highlights, color shift due to chrome reflectance, and missing absorption in renderings (Figs. 16–19). Quantitative comparisons confirmed improvement due to perspective unwrapping of light probe images and led to analysis by synthesis.

The comparison with reference photographs before and after deformation (Fig. 17) to some extent validates our soft object deformation technique. Further validation would be desirable, but it is difficult to come up with a different experiment. Some kind of soft, durable memory foam with a scannable surface would be required as the soft object would otherwise change shape again once the hard object is removed. Our validation only supports that the cloth appearance (as observed through glass) is represented more faithfully after deformation.

We found analysis by synthesis useful for estimating parameters with an outset in physics-based initial guesses. The results in Fig. 12 show that we can estimate optical properties for a given material and use them in a different setting (right part of Fig. 1). The precision of the estimation varies with the impact of the property on the overall error, and the estimated parameters may compensate for unrelated errors. In this regard, specific scene configurations could be used to favor estimation of a particular parameter.

The most important limitation of our method is that we describe materials as large patches of isotropic BRDFs. In our renderings, this assumptions works well for the checkerboard backdrop but not for the cloth, where we both have subsurface scattering effects and probably anisotropy due to the weave structure of the cloth. Fig. 21 reveals that the rendered image is too dark in areas surrounding caustics. As seen in the light refracted through the sphere in the vicinity of the marker, our processing of the glass object to separate glass from markers causes some imprecision in the geometry. We believe this mainly influences the shape of the caustic. The bleeding of the caustic to areas that are much darker in the rendered images looks like backscattering from the table beneath the cloth. We refer to this as a kind of subsurface scattering.

Another limitation is seen at the transition between non-connected elements. It is visible in the renderings at the boundary between the cloth and the backdrop (see Fig. 20). The problem derives from the fact that the cloth and the backdrop were too close to each other during dataset acquisition. This resulted in the Poisson mesh reconstruction interpreting them as a continuous object instead of two separate ones. The problems around markers (Fig. 21) are also due to transition of materials. The marker removal and whole closing in the glass surface reconstruction interrupts the original shape of the surface. Furthermore, the markers are glued onto the glass surface, and the glue is not considered in the reconstruction and renderings. The marker glue problem is magnified by the glass refraction.

6. CONCLUSION

We have proposed a pipeline for multimodal scene digitization. Our work addresses the entire process from acquisition of the original objects, through reassembly of the digital scene, to accurate modeling of camera and environment. While the pipeline required several non-trivial steps, the benefits are correspondingly great since we can perform pixelwise comparisons between rendered images and photographs of the corresponding physical scene. This means that we have the means to quantitatively assess the accuracy of an acquired model based on comparison with empirical evidence. We believe this kind of quantitative assessment has not previously been possible for transparent objects. In applications like cultural heritage preservation and industrial inspection, where the accuracy of a digitization is important, such comparison with empirical evidence is crucial.

To the best of our knowledge, our work is also the first work to quantify the photorealism of a heterogeneous scene requiring multimodal acquisition.

Our dataset is publicly available so that others can test new techniques for the different steps of the pipeline with quantitative feedback based on photorealistic rendering. The fact that one can use off-the-shelf rendering techniques for improving the different steps of a multimodal digitization pipeline is perhaps the most important benefit of our work. An application of the full pipeline is the virtual product placement in Fig. 1. Another important application is the estimation of radiometric properties through analysis by synthesis. The ability to accurately estimate optical properties through computation rather than measurement, which might require specialized equipment, is likely to greatly simplify the digitization of radiometrically complex objects. In this paper, we estimated absorption and refractive indices of transparent objects, but analysis by synthesis could be equally useful for other materials with non-trivial BRDFs. This is another key benefit of our work that we believe is well worth exploring in the future.

Fig. 22. Comparison of renderings and photographs as in Fig. 1 (left), but with more views.

Funding. Innovation Fund Denmark (IFD) (75-2014-1, 3067-00001B, 5163-00001B, 5163-00003B).

A. APPENDIX

Figure 22.

REFERENCES

1. M. Weinmann and R. Klein, "Advances in geometry and reflectance acquisition (course notes)," in "Proceedings of SIGGRAPH Asia 2015 Courses," (ACM, 2015).
2. P. Debevec, "The light stages and their applications to photoreal digitial actors," in "SIGGRAPH Asia 2012 Technical Briefs," (2012).
3. L. Gomes, O. R. P. Bellon, and L. Silva, "3D reconstruction methods for digital preservation of cultural heritage: A survey," Pattern Recognition Letters **50**, 3–14 (2014).
4. L. Zhang, H. Dong, and A. E. Saddik, "From 3D sensing to printing: A survey," ACM Transactions on Multimedia Computing, Communications, and Applications **12**, 27:1–27:23 (2016).
5. J. B. Nielsen, E. R. Eiriksson, R. L. Kristensen, J. Wilm, J. R. Frisvad, K. Conradsen, and H. Aanæs, "Quality assurance based on descriptive and parsimonious appearance models," in "Workshop on Material Appearance Modeling (MAM 2015)," (The Eurographics Association, 2015), pp. 21–24.
6. B. T. Phong, "Illumination for computer generated pictures," Communications of the ACM **18**, 311–317 (1975).
7. C. M. Goral, K. E. Torrance, D. P. Greenberg, and B. Battaile, "Modeling the interaction of light between diffuse surfaces," Computer Graphics (Proceedings of SIGGRAPH 84) **18**, 213–222 (1984).
8. A. Takagi, H. Takaoka, T. Oshima, and Y. Ogata, "Accurate rendering technique based on colorimetric conception," Computer Graphics (Proceedings of SIGGRAPH 90) **24**, 263–272 (1990).
9. G. W. Meyer, H. E. Rushmeier, M. F. Cohen, D. P. Greenberg, and K. E. Torrance, "An experimental evaluation of computer graphics imagery," ACM Transactions on Graphics **5**, 30–50 (1986).
10. H. Rushmeier, G. Ward, C. Piatko, P. Sanders, and B. Rust, "Comparing real and synthetic images: Some ideas about metrics," in "Rendering Techniques '95 (Proceedings of EGWR 1995)," (Springer, 1995), pp. 82–91.
11. K. F. Karner and M. Prantl, "A concept for evaluating the accuracy of computer generated images," in "Proceedings of Spring Conference on Computer Graphics (SCCG 1996)," (1996).
12. S. N. Pattanaik, J. A. Ferwerda, K. E. Torrance, and D. P. Greenberg, "Validation of global illumination solutions through CCD camera measurements," in "Proceedings of Color Imaging Conference (CIC 1997)," (1997), pp. 250–253.
13. N. L. Jones and C. F. Reinhart, "Parallel multiple-bounce irradiance caching," Computer Graphics Forum (Proceedings of EGSR 2016) **35**, 57–66 (2016).
14. N. L. Jones and C. F. Reinhart, "Experimental validation of ray tracing as a means of image-based visual discomfort prediction," Building and Environment **113**, 131–150 (2017).
15. D. P. Greenberg, K. E. Torrance, P. Shirley, J. Arvo, J. A.Ferwerda, S. Pattanaik, E. Lafortune, B. Walter, S.-C. Foo, and B. Trumbore, "A framework for realistic image synthesis," in "Proceedings of SIGGRAPH 97," (ACM/Addison-Wesley, 1997), pp. 477–494.
16. F. Drago and K. Myszkowski, "Validation proposal for global illumination and rendering techniques," Computers & Graphics **25**, 511–518 (2001).
17. C. Ulbricht, A. Wilkie, and W. Purgathofer, "Verification of physically based rendering algorithms," Computer Graphics Forum **25**, 237–255 (2006).
18. J. Meseth, G. Müller, R. Klein, F. Röder, and M. Arnold, "Verification of rendering quality from measured BTFs," in "Proceedings of Applied Perception in Graphics and Visualization (APGV 2006)," (ACM, 2006), pp. 127–134.
19. A. I. Ruppertsberg and M. Bloj, "Rendering complex scenes for psychophysics using RADIANCE: How accurate can you get?" Journal of the Optical Society of America A **23**, 759–768 (2006).
20. A. Dal Corso, J. R. Frisvad, T. K. Kjeldsen, and J. A. Bærentzen, "Interactive appearance prediction for cloudy beverages," in "Workshop on Material Appearance Modeling (MAM 2016)," (The Eurographics Association, 2016), pp. 1–4.
21. B. Tunwattanapong, G. Fyffe, P. Graham, J. Busch, X. Yu, A. Ghosh, and P. Debevec, "Acquiring reflectance and shape from continuous spherical harmonic illumination," ACM Transactions on Graphics (Proceedings of SIGGRAPH 2013) **32**, 109:1–109:11 (2013).
22. T. Nöll, J. Köhler, G. Reis, and D. Stricker, "Fully automatic, omnidirectional acquisition of geometry and appearance in the context of cultural heritage preservation," Journal on Computing and Cultural Heritage **8**, Article 2 (2015).
23. H. Wu, Z. Wang, and K. Zhou, "Simultaneous localization and appearance estimation with a consumer RGB-D camera," IEEE Transactions on Visualization and Computer Graphics **22**, 2012–2023 (2016).
24. I. Ihrke, K. N. Kutulakos, H. P. A. Lensch, M. Magnor, and W. Heidrich, "Transparent and specular object reconstruction," Computer Graphics Forum **29**, 2400–2426 (2010).
25. A. Kolb, J. Zhu, and R. Yang, "Sensor fusion," in "Digital Representation of the Real World," M. A. Magnor, O. Grau, O. Sorkine-Hornung, and C. Theobalt, eds. (CRC Press, 2015), chap. 9, pp. 133–150.
26. V. Bhateja, H. Patel, A. Krishn, A. Sahu, and A. Lay-Ekuakille, "Multimodal medical image sensor fusion framework using cascade of wavelet and contourlet transform domains," IEEE Sensors Journal **15**, 6783–6790 (2015).
27. A. Pamart, O. Guillon, J.-M. Vallet, and L. De Luca, "Toward a multimodal photogrammetric acquisition and processing methodology for monitoring conservation and restoration studies," in "Eurographics Workshop on Graphics and Cultural Heritage," (The Eurographics Association, 2016), pp. 207–210.
28. H. Aanæs and A. B. Dahl, "Accuracy in robot generated image data sets," in "Proceedings of SCIA 2015," , vol. 9127 of *Lecture Notes in Computer Science* (Springer, 2015), pp. 472–479.
29. H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," International Journal of Computer Vision **120**, 153–168 (2016).
30. Z. Zhang, "A flexible new technique for camera calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 1330–1334 (2000).
31. P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in "Proceedings of SIGGRAPH 97," (ACM/Addison-Wesley, 1997), pp. 369–378.
32. J. L. Posdamer and M. Altschuler, "Surface measurement by space-encoded projected beam systems," Computer Graphics and Image Processing **18**, 1–17 (1982).
33. J. Geng, "Structured-light 3D surface imaging: a tutorial," Advances in

Optics and Photonics **3**, 128–160 (2011).
34. M. Kazhdan and H. Hoppe, "Screened Poisson surface reconstruction," ACM Transactions on Graphics **32**, 29:1–29:13 (2013).
35. M. Corsini, P. Cignoni, and R. Scopigno, "Efficient and flexible sampling with blue noise properties of triangular meshes," IEEE Transactions on Visualization and Computer Graphics **18**, 914–924 (2012).
36. J. F. Murray-Coleman and A. M. Smith, "The automated measurement of BRDFs and their application to luminaire modeling," Journal of the Illuminating Engineering Society **19**, 87–99 (1990).
37. J. B. Nielsen, H. W. Jensen, and R. Ramamoorthi, "On optimal, minimal BRDF sampling for reflectance acquisition," ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2015) **34**, 186:1–186:11 (2015).
38. W. Matusik, H. Pfister, M. Brand, and L. McMillan, "A data-driven reflectance model," ACM Transactions on Graphics (Proceedings of SIGGRAPH 2003) **22**, 759–769 (2003).
39. J. F. Barrett and N. Keat, "Artifacts in CT: Recognition and avoidance," RadioGraphics **24**, 1679–1691 (2004).
40. T. Ju, F. Losasso, S. Schaefer, and J. Warren, "Dual contouring of Hermite data," ACM Transactions Graphics (Proceedings of SIGGRAPH 2002) **21**, 339–346 (2002).
41. M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM **24**, 381–395 (1981).
42. L. Kobbelt, "$\sqrt{3}$-subdivision," in "Proceedings of SIGGRAPH 2000," (ACM/Addison-Wesley, 2000), pp. 103–112.
43. R. L. Cook, "The Reyes image rendering architecture," Computer Graphics (Proceedings of SIGGRAPH 87) **21**, 95–102 (1987).
44. T. Atherton and D. Kerbyson, "Size invariant circle detection," Image and Vision Computing **17**, 795 – 803 (1999).
45. P. D. Sampson, "Fitting conic sections to "very scattered" data: An iterative refinement of the Bookstein algorithm," Computer Graphics and Image Processing **18**, 97–108 (1982).
46. J. C. Gower, "Generalized Procrustes analysis," Psychometrika **40**, 33–51 (1975).
47. G. D. Finlayson, M. Mackiewicz, and A. Hurlbert, "Color correction using root-polynomial regression," IEEE Transactions on Image Processing **24**, 1460–1470 (2015).
48. G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," Color Research & Application **30**, 21–30 (2005).
49. J. Nocedal and S. J. Wright, *Numerical Optimization* (Springer, 2006), 2nd ed.
50. E. Reinhard, G. Ward, S. Pattanaik, P. Debevec, W. Heidrich, and K. Myszkowski, *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting* (Morgan Kaufmann/Elsevier, 2010), 2nd ed.
51. A. Stockman and L. T. Sharpe, "The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype," Vision Research **40**, 1711–1737 (2000).
52. J. R. Frisvad, N. J. Christensen, and H. W. Jensen, "Computing the scattering properties of participating media using Lorenz-Mie theory," ACM Transactions on Graphics (Proceedings of SIGGRAPH 2007) **26**, 60:1–60:10 (2007).
53. C. Ulbricht and A. Wilkie, "A problem with the use of XYZ colour space for photorealistic rendering computations," in "Proceedings of Colour in Graphics, Imaging, and Vision (CGIV 2006)," (2006), pp. 435–437.
54. J. Meng, F. Simon, J. Hanika, and C. Dachsbacher, "Physically meaningful rendering using tristimulus colours," Computer Graphics Forum (Proceedings of EGSR 2015) **34**, 31–40 (2015).
55. H. S. Fairman, M. H. Brill, and H. Hemmendinger, "How the CIE 1931 color-matching functions were derived from Wright-Guild data," Color Research & Application **22**, 11–23 (1997).
56. P. Debevec, "Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography," in "Proceedings of SIGGRAPH 98," (ACM, 1998), pp. 189–198.
57. A. D. Rakić, A. B. Djurišić, J. M. Elazar, and M. L. Majewski, "Optical properties of metallic films for vertical-cavity optoelectronic devices," Applied Optics **37**, 5271–5283 (1998).
58. J. T. Kajiya, "The rendering equation," Computer Graphics (Proceedings of SIGGRAPH 86) **20**, 143–150 (1986).
59. M. Pharr, W. Jakob, and G. Humphreys, *Physically Based Rendering: From Theory to Implementation* (Morgan Kaufmann/Elsevier, 2017), 3rd ed.
60. S. G. Parker, J. Bigler, A. Dietrich, H. Friedrich, J. Hoberock, D. Luebke, D. McAllister, M. McGuire, K. Morley, A. Robison, and M. Stich, "OptiX: A general purpose ray tracing engine," ACM Transactions on Graphics (Proceedings of SIGGRAPH 2010) **29**, 66:1–66:13 (2010).
61. M. Hejrati and D. Ramanan, "Analysis by synthesis: 3D object recognition by object reconstruction," in "Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)," (2014), pp. 2449–2456.
62. D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," Journal of Global Optimization **13**, 455–492 (1998).
63. G. Taubin, "A signal processing approach to fair surface design," in "Proceedings of SIGGRAPH 95," (ACM Press, 1995), pp. 351–358.

Contribution B

A variational study on BRDF reconstruction in a structured light scanner

A Variational Study on BRDF Reconstruction in a Structured Light Scanner

Jannik Boll Nielsen, Jonathan Dyssel Stets, Rasmus Ahrenkiel Lyngby,
Henrik Aanæs, Anders Bjorholm Dahl, Jeppe Revall Frisvad
Technical University of Denmark
http://eco3d.compute.dtu.dk/

Abstract

Time-efficient acquisition of reflectance behavior together with surface geometry is a challenging problem. In this study, we investigate the impact of system parameter uncertainties when incorporating a data-driven BRDF reconstruction approach into the standard pipeline of a structured light scanning system. The parameters investigated include geometric detail of scanned objects; vertex positions and normals; and position and intensity of light sources. To have full control of uncertainties, experiments are carried out in a simulated environment, mimicking an actual structured light scanning setup. Results show that while uncertainties in vertex positions and normals have a high impact on the quality of reconstructed BRDFs, object geometry and light source properties have very little influence on the reconstructed BRDFs. With this analysis, practitioners now have insight in the tolerances required for accurate BRDF acquisition to work.

1. Introduction

The topic of accurate appearance capture and digitization is gaining attention in areas like the movie and gaming industries [9], preservation of cultural heritage [6], and quality assurance in production [18]. These applications demand automatic and fast systems that can acquire full and accurate appearance, including both radiometry and geometry. In combination, these two components define appearance, and numerous methods have been proposed for their acquisition. Capturing high quality geometric models of real world objects is today a well-addressed problem with many good solutions. Different technologies exist such as structured light (SL) scanners, multi-view stereo, or time-of-flight, each having their own advantages and disadvantages. With respect to radiometric properties, techniques such as goniometric setups, curved mirror configurations, and light domes can be used for accurately estimating bidirectional reflectance distribution functions (BRDFs) of simple, often flat, geometries. However, robust approaches for jointly estimating radiometry and geometry are few and often require advanced and expensive setups or produce low quality results.

In this paper we investigate how a SL scanner, designed for high quality geometry acquisition, can be modified with few adjustments to also capture reflectance samples. Thus, the scanner can also sample the BRDF of a scanned object and reconstruct it using state of the art BRDF reconstruction methods. Using this system as an offset, we investigate the influence on BRDF estimation caused by various system uncertainties. The uncertainties investigated include: geometric complexity of the scanned object, vertex position and normal, and light source position and intensity. Our aim is to gain insight into how BRDF reconstruction is affected by various error sources and uncertainties. As a main result, we provide a lookup table for system designers, telling them the system specifications required for correctly estimating BRDFs in a given material/geometry configuration. In order to ensure full control of all uncertainties, the experiment is designed as a simulation of an SL scanner system. The simulation is based on real world parameters from an actual SL scanning system, as well as real measured BRDFs from the MERL database [17].

Although this study focuses on an SL scanning system, we believe that the proposed modification, as well as the insights into the influence of error sources, applies to most 3D scanning systems where an image-forming sensor and a light source is present. Likewise, while we apply the BRDF reconstruction framework of Nielsen et al. [20], we expect other BRDF modeling frameworks with strong priors to be applicable as well.

2. Related Work

A multitude of techniques exist for acquiring shape and appearance [30]. Most techniques are time consuming or require highly specialized equipment. In the following, we relate our work to instrumental setups that are similar to the one we propose. Our setup is a structured light 3D scanner setup with two cameras, a projector light source, and a turntable. An additional LED source is added to our setup.

An example of early work investigating the acquisition of shape and reflectance properties using images is that of Ikeuchi and Sato [11]. They fit the Torrance-Sparrow BRDF model [28] to samples obtained from a range image and a brightness image. To investigate the convergence of their method to true values (robustness), they do a simulation study based on rendered images with different noise levels applied. This enables them to draw important conclusions with respect to the sensitivity and range of applicability of their method. Unfortunately, it seems that such simulation studies are very uncommon in subsequent work in this area. To fill this gap, we present a simulation study of this kind for our more contemporary acquisition technique.

The idea of a camera, a light source, and a turntable for joint acquisition of shape and appearance (surface geometry and BRDF) was pioneered by Lu and Little [15]. They use a collimated source and estimate the BRDF for (near) zero half-angle by finding the points of maximum intensity and tracking them as the object turns around its axis. After this, they acquire the surface geometry using a shape from shading approach. Their approach requires assumption of a smooth object and a uniform BRDF across the object surface. The instrument we consider is similar in complexity, but based on a structured light setup with a projector light source and two cameras (stereo). We also flip the procedure and acquire shape using structured light, and then we estimate a full isotropic BRDF.

It is interesting to note that Lu and Little [15] try perturbations of depth and rotation axis to investigate robustness of their technique. In addition, they indicate that experiments on synthetic images to perform a more in-depth investigation would be appropriate. Nevertheless, we are unable to find such an investigation in the work following that of Lu and Little. Our goal is thus to provide one.

Based on robot arm sample rotation and a structured light range scanner, Sato and Ikeuchi [24] extend their earlier (range and brightness image) technique to include scan of the full geometry of an object and estimation of its spatially varying reflectance properties. The reflectance properties are, however, parameters in an analytic BRDF model and no BRDF ground truth is available for validation. Marschner et al. [16] propose a similar technique, but based on a handheld camera and the Lafortune BRDF model [13]. Employing a more conventional structured light 3D scanner (or a computed tomography scanner) to obtain surface geometry, Lensch et al. [14] extend the technique to acquire Lafortune model parameters for spatially varying BRDFs.

Krzesłowski et al. [12] present a structured light scanner with added LED sources for integrated acquisition of BRDF and surface geometry. However, they fit their sampled BRDF data to the Blinn-Phong model [2, 22], which only provides a good BRDF fit for a limited range of materials [19]. The structured light scan provides a sparse sampling of the BRDF per sample point in the scanned surface geometry. The Blinn-Phong model is fitted to this sparse set of BRDF samples. The acquisition approach we investigate is similar, but we do a simulation study to identify the impact of different potential error sources. We limit our study to an object with just one BRDF across the object surface, and we use the BRDF model of Nielsen et al. [20].

Using a beam splitter to have coaxial camera and projector light source, Holroyd et al. [10] develop a gonioreflectometer which can also acquire the surface geometry using structured light. While this technique delivers high quality acquisitions, it is not a time-efficient approach like a structured light setup. Sitnik et al. [27] propose a faster integrated measurement system with a single image sensor. Here, a multi-spectral camera is combined with a projector and a grid of 16 broadband light sources to capture both the 3D geometry and multi-spectral light intensity information. In another complex setup, Tunwattanapong et al. [29] propose a rotating light arc providing spherical harmonic illumination used together with five cameras to reconstruct reflectance maps. The geometry is then reconstructed using multi-view stereo based on the diffuse and specular reflectance maps. Finally, Schwartz et al. [25] propose a system, based on SL and HDR imaging, for measuring bidirectional texure functions (BTFs) using a light dome composed of 188 LEDs, four projectors, eleven cameras and a rotation stage. The complexity of these instrumental setups is significantly higher than the SL setup that we propose.

3. Implementation

In this study, the BRDF estimation process revolves around a structured light scanning system like the one illustrated in figure 1. The system is composed of two cameras used for triangulation, a projector for projecting an encoding pattern, a rotation stage for rotating a sample, and a scene light. The principles behind the approach should be applicable to any 3D scanning system comprised of components including an image-forming sensor and a light source. In the following subsections, the modified SL capturing pipeline is outlined along with the reconstruction method. The implementation of the modifications required for a structured light scanning system to estimate BRDFs is fairly straightforward in practice, however, to ensure full control of all variables in the study, as well as avoiding unforeseen noise sources, the reflectance acquisition part of the pipeline is here simulated. Below, the details of this simulation process will also be covered.

3.1. Capture Pipeline

The principles behind estimating a BRDF in the SL pipeline are based on the assumption that the BRDF can be observed under a sufficient number of view/light configurations. We need enough to confidently fit a model to the ob-

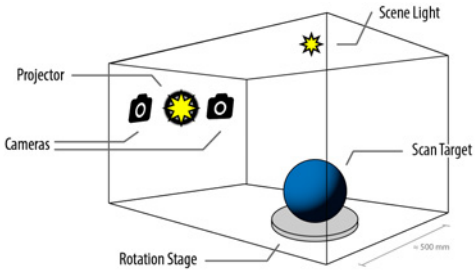


Figure 1: Structured light scanning system consisting of two cameras used for triangulation, a projector for projecting an encoded pattern, a rotation stage for rotating the sample, and a scene light.

servations. Enough configurations are obtained in scenarios where a scanned object, with a sufficiently varying surface geometry, consists of a homogeneous material and is rotated during the capturing process. Any point on the surface will thus be observed under different view/light configurations, and with a sufficiently large number of points with unique surface normals, a sufficiently large number of BRDF samples can be acquired for reconstruction.

Clearly, the full four-dimensional space spanned by the BRDF will not be covered by these observations, let alone due to the fixed baseline between light source (projector) and observer (cameras), which corresponds to a fixed difference angle (θ_d) in the Rusinkiewicz parametrization [23]. Even in a better posed scenario as figure 1, where an additional scene light is present, the BRDF space is still very sparsely sampled. Nonetheless, a sufficient number of observations can in fact be acquired through this process if we use a strong prior when fitting a BRDF model.

The SL scanning pipeline involves projecting an encoding pattern onto the target object and triangulating the encoded pixels seen by the camera(s). This is sometimes followed by a rotation of the sample, after which the scanning is repeated. The modification to the standard SL scanning pipeline is simple and consists only in capturing a high dynamic range (HDR) image of the sample. This is done before the sample is rotated (or removed) using the triangulation camera(s) and a fully lit projector. If a scene light is present, as it is here, an additional HDR image is captured under its illumination. With the captured HDR images, it is possible in post-processing to reproject the captured vertices onto these and acquire a radiance value. With knowledge of vertex normal, camera position, light source position, and light source intensity, this radiance value may be converted into a BRDF sample, defined by

$$f_r\left(\boldsymbol{\omega}_i, \boldsymbol{\omega}_r\right) = \frac{dL_r\left(\boldsymbol{\omega}_r\right)}{dE_i\left(\boldsymbol{\omega}_i\right)}, \tag{1}$$

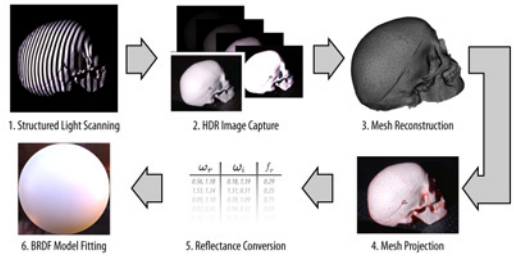


Figure 2: Geometry and BRDF capture pipeline in a structured light scanning system.

which is the ratio between the radiance reflected off a surface in a specific direction and the irradiance hitting a surface from another specific direction.

The overall capturing pipeline is depicted in figure 2. The pipeline consists of:

1. Structured light scanning
2. HDR image capture
3. Mesh reconstruction
4. Mesh projection onto HDR images
5. Per vertex HDR intensity to reflectance conversion
6. BRDF model fitting

In step 1, a traditional 3D scanning is carried out, in this case using structured light. Before altering anything in the scene in any way, e.g. by rotating or removing the sample, an HDR image is captured in step 2 using the multiple exposure approach of Debevec et al. [3]. This ensures a floating point precision image conforming with the scanned geometry and camera calibration of the SL scanner. The acquisition part is followed by post processing, initialized with a meshing in step 3 of the acquired point cloud. As will later become apparent, a mesh is required for filtering purposes. In step 4, the mesh is projected onto the HDR images, assigning every vertex with an HDR intensity. All vertex intensities are in step 5 converted to reflectance values based on scene geometry, and finally in step 6 a BRDF model is fitted to the observed BRDF samples.

3.1.1 Structured Light Scanner

In order to provide a thorough description of our method, we briefly outline our SL scanning strategy. Please note that this is by no means a complete description. For specific details, we refer to the work of others [8, 33, 4].

SL scanning is a form of stereo vision. Essentially, stereo vision is the process of reconstructing the 3D shape of an object by using a set of cameras as protractors. The pixel positions, and thereby the incident angles, of a given 3D

point are found in the camera images. From knowing the mutual transformations between the cameras, the 3D position of the point can be computed based on trigonometry. The key difficulty is finding corresponding points in the images. SL based techniques seek to lower the complexity of this correspondence problem by projecting a known pattern onto the reconstruction object. There are a plethora of encoding strategies available [5], but they all seek to assign unique ID numbers to pixels based on their distance from the projector. These ID numbers are then used to determine pixel correspondences, and from that compute the depth of the surface under the pixels.

Based on the conclusions made by Eiríksson et al. [4], we have selected a scanner system composed of two cameras and one projector which uses the phase shifting (PS) encoding strategy [7]. In short, the projector projects a series of spatially distributed gray-scale sinusoidal patterns onto the target surface. Each pattern has a given frequency and phase shift. We use three frequencies with up to 32 phase shifts per frequency for a total of 64 patterns.

3.1.2 Vertex Reflectance Assignment

From the calibration of the SL scanning system, the intrinsics and extrinsics have been determined. Commonly these are described by a pinhole camera model with a projection matrix $\boldsymbol{P}$ given as:

$$\boldsymbol{P} = \boldsymbol{K}\left[\boldsymbol{R}\,\boldsymbol{t}\right], \tag{2}$$

with $\boldsymbol{R}$ and $\boldsymbol{t}$ being the rotation and translation of the camera respectively, and $\boldsymbol{K}$ being the intrinsic parameters of the camera [34]. With this, any 3D point in homogeneous coordinates, $\boldsymbol{q}$, may be projected onto the cameras 2D image plane by:

$$\hat{\boldsymbol{q}} = \boldsymbol{P}\boldsymbol{q}. \tag{3}$$

Thus, any vertex from a scanned object may be reprojected onto its corresponding HDR image and have a specific radiance RGB value assigned to it. By calibration with e.g. Spectralon, the light intensity at the sample can be predetermined, and often this intensity can be assumed constant over the physical span of the sample. With this prior knowledge, and correcting with the cosine between light and vertex normal, the vertex radiance value may be converted into a BRDF value:

$$f_r = \frac{\texttt{HDR}(\boldsymbol{P}\boldsymbol{v}_{\texttt{position}})}{(\boldsymbol{\omega}_i \cdot \boldsymbol{v}_{\texttt{normal}})\, I}, \tag{4}$$

where $\texttt{HDR}(\hat{\boldsymbol{q}})$ is the HDR radiance value at position $\hat{\boldsymbol{q}}$, $\boldsymbol{v}$ is the vertex, $\boldsymbol{\omega}_i$ is the normalized light direction, and I is the predetermined light intensity at the position of the scanned sample.

Note that some vertices may be projected into shadow regions in the HDR image. In order to avoid this, two tests are employed. First, all vertices with a normal facing away from the camera or light are removed, this is the case when $\boldsymbol{\omega}_{r/i} \cdot \boldsymbol{v}_{\texttt{normal}} <= 0$. This test filters most invalid observations away, but in scenarios where self-shadowing may occur, a shadow map calculation is also applied [31]. This, however, requires that the scanned object has been converted into a 3D mesh, which in itself may introduce artifacts if care is not taken.

3.1.3 Data-Driven BRDF Reconstruction

The challenge of fitting a reflectance model to the sparse number of BRDF samples calls for a model with a strong prior. In this study, the data-driven BRDF reconstruction framework of Nielsen et al. [20, 32] is chosen for this purpose, as it is known to work well for problems where only very few BRDF samples are available. The model is based on the MERL database [17] of isotropic BRDFs spanning a wide range of common materials. Using a log-relative mapping of reflectance values, projections in principal component space allows inferring missing observations from existing ones. Effectively the model reconstructs a MERL format BRDF, i.e. a $90 \times 90 \times 180$ bin tabulated isotropic BRDF, from any number of input observations provided. The biggest limitation of this approach is that it requires the measured material to lie within the convex hull spanned by the MERL database. If this is met, under ideal lighting conditions, as little as two images are sufficient to faithfully reproduce a material.

3.2. Simulation of Pipeline

In order to maintain full control of all uncertainties in this fairly complex acquisition pipeline, a simulated pipeline is used to produce realistic HDR images, conforming with a true SL system. We do this by initially picking a ground truth mesh and ground truth measured BRDF from the MERL database. Using these, combined with the true SL system projection matrices, light source positions, and rotation stage positions, an OpenGL renderer is used to produce a series of HDR renderings of the chosen geometry and BRDF as it would have been seen by the SL system. An example of such renderings is shown in figure 3, where 3 different meshes with the "blue-rubber" BRDF applied have been rendered as would be seen by the SL scanning system (although cropped here). With this, the ground truth appearance behind every HDR image is available, allowing for a quantitative evaluation of reconstruction.

3.2.1 Dataset Generation

Four different types of materials and three different types of geometries were chosen to generate the evaluated dataset. Material-wise, four different levels of specularity were chosen, all in different colors, covering the span of material

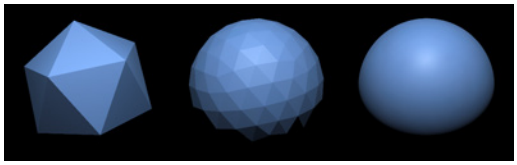


Figure 3: Icospheres with 3 different tessellation levels: 1, 3 and 5. For the highest tesselation level vertex normals have been smoothed.

behavior that would be expected in the real world. The materials are "blue-rubber", "green-metallic-paint", "purple-plastic", and "specular-black-phenolic", with the first having very soft highlights and the last being highly specular (renderings are available in the lower left corner of figures 5–8). As the data-driven BRDF reconstruction model is also based on the MERL database, these four materials were excluded in the model-training. Geometry-wise three different geometries were chosen, spanning the amount of geometric detail that can be expected from real world objects. The geometries are based on an icosphere with increasing tesselation levels and are shown in figure 3. This is motivated by the fact that a sphere naturally covers all possible surface normals, while a plane only covers a single. Thus, the closer the geometry is to a sphere, the more ideal are the BRDF reconstruction conditions from a geometry point of view. Each of the three meshes has been subdivided to consist of roughly 15000 vertices and are scaled to have a diameter of 100 mm in the simulator.

We generate a dataset of HDR images using the materials and geometries described above. Both the scene-light and projector are used as light sources and both cameras are used for observing, see figure 1. In addition, the sample is rotated in 10 steps from 0° to 180°. This gives $n_{\texttt{conf}} = n_{\texttt{rot}} \times n_{\texttt{lights}} \times n_{\texttt{cameras}} = 10 \times 2 \times 2 = 40$ HDR images per material/geometry configuration and $n_{\texttt{total}} = n_{\texttt{conf}} \times n_{\texttt{materials}} \times = n_{\texttt{geometries}} = 40 \times 4 \times 3 = 480$ HDR images in total.

3.2.2 Noise Addition

There are a range of elements in the pipeline depicted in figure 2 that affect the accuracy of the BRDF observations acquired. Any uncertainties in these will obviously cause uncertainties in the BRDF model-fitting. To gain insight into this, four types of uncertainties are investigated:

Vertex position. The precision of the SL system will determine the geometric noise present in a 3D scan. Clearly, as the vertices are projected onto HDR images, any error in position will cause a wrong assignment of radiance value.

Vertex normal. Commonly, surface normals are not a direct product of the 3D acquisition procedure but are estimated afterwards, e.g. based on the spatial distribution of neighboring vertices. This makes the estimation prone to errors, and any wrong orientation of normals will directly influence the reflectance estimate.

Light position. While camera positions are very precisely calibrated, the light position is oftentimes significantly more difficult to determine. The position affects the light direction and thus also the reflectance estimate.

Light intensity. Finally, precise knowledge of the light intensity at any given 3D point in the SL system is not easily obtained. As the light intensity is used to compute the fraction of light reflected off the material surface, it too directly influences the reflectance estimate.

As the evaluated dataset is simulated, the exact system parameters are known. This allows for, prior to processing the data, manually adding a controlled amount of noise to any of the above components. To apply noise in our experiments, we use a normal distribution (Gaussian noise) with the given position or normal as mean and σ is standard deviation. For normals, the noise only applies to the polar angle. To add noise in the case of light intensity, we multiply the intensity by a normal distribution with unit mean and $\sigma/100$ as standard deviation (percentage noise).

3.2.3 Evaluation

Evaluating the quality of an estimated BRDF compared to the ground truth is not trivial and is indeed a research field in itself. In these experiments, both qualitative and quantitative measures are presented:

In-plane reflectance profiles. For qualitative evaluation, 45° in-plane reflectance profiles of estimated and ground truth BRDFs are presented. These plots visualize the general shape of the specular highlight as well as parts of the grazing angle behaviour.

Ray-traced sphere renderings. Another qualitative evaluation is using a physically based renderer [21]. Here the BRDFs can be visualized under realistic environment lighting conditions, giving the viewer an impression of how the material would look in the real world. The material examples shown in figures 5–8 are rendered this way.

Tone mapped color difference. Rendered images, using the approach above, of the ground truth and reconstructed BRDFs are compared using the CIEDE2000 color difference perception measure. The HDR images are first scaled to the visible range using Reinhard tonemapping, and gamma correction ($\gamma = 2.2$) at F-stop 0 is applied [1]. The images are then converted to the CIE 1976 L*a*b* color space, and the CIEDE2000 color difference formula [26] (with $[k_L\ k_C\ k_H] = [1\ 1\ 1]$) is used to calculate the color difference ΔE_{00}. The average of all pixel differences is calculated and used as a perceptual similarity measure between

	blue-rubber	green-metallic-paint	purple-paint	specular-black-phenolic
Icosphere 1	0.77 ± 1.02	2.77 ± 2.93	1.56 ± 1.96	1.07 ± 1.13
Icosphere 3	0.37 ± 0.78	2.60 ± 3.11	0.82 ± 1.01	2.50 ± 3.14
Icosphere 5	0.41 ± 0.67	3.00 ± 3.23	0.55 ± 0.75	1.43 ± 1.92
Icosphere 5*	0.52 ± 0.96	5.19 ± 5.20	1.58 ± 1.55	2.29 ± 1.63

Table 1: Errors for increasing geometric detail (icosphere tesselation level). Errors are measured as the average ΔE_{00} color difference between tone mapped renderings of ground truth BRDF and reconstruction. Icosphere 1,3,5 are reconstructions using two light sources, while 5^* are reconstructions using only the projector as light source.

images, and the standard deviation represents the certainty of this number.

4. Results

We report results for BRDF estimation under various noise influences. This includes an evaluation of BRDF estimation performance under three different geometry complexities, followed by an evaluation of performance under influence of uncertainties with respect to vertex position, vertex normal, light source position, and light source intensity. Due to page limitations, some comparisons of in-plane reflectance profiles and renderings have been omitted. A summary of comparisons are reported in tables 1 and 2.

4.1. Geometry Dependency

In order to evaluate how much geometric complexity affects the quality of an estimated BRDF, estimations were carried out on the simulated icospheres with tesselation levels 1, 3 and 5, depicted in figure 3. The estimates were computed under ideal conditions, i.e. no noise added to any of the system components listed in section 3.2.2. In figure 4, quantitative comparisons of the material "purple-paint" are presented in the form of in-plane reflectance profiles and renderings. It may be seen that as geometric detail increases, the quality of reconstruction improves, however the improvement is surprisingly small. In table 1, the results for all four materials are listed, using the ΔE_{00} color-difference measure between ground truth rendering and reconstructed rendering. To the convenience of system designers, errors using icosphere level 5 combined with only the projector as light source is also presented in the bottom row of table 1.

To provide as ideal conditions as possible for the noise simulations, the icosphere level 5 geometry will be used in the following evaluations. For all evaluations, 30 repetitions were carried out to estimate mean and standard deviation of reconstruction. Quantitative comparisons for all materials,

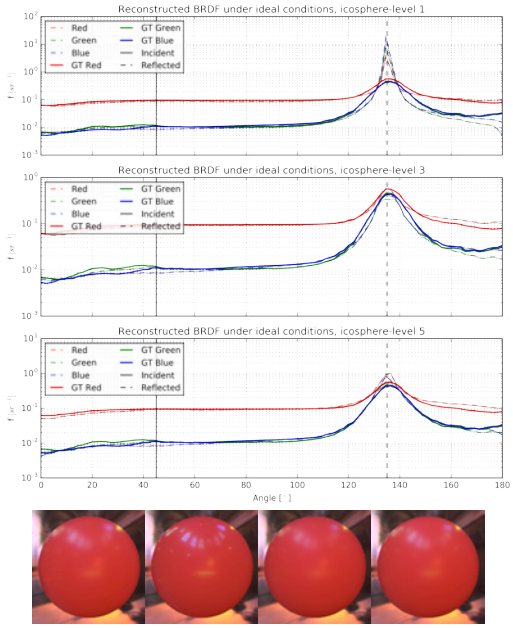


Figure 4: Ideal reconstructions of "purple-paint" material, using icosphere tesselation levels $\{1, 3, 5\}$, shown as $45°$ in-plane profiles. Solid lines indicate ground truth BRDF RGB channels, dashed lines are the reconstructed BRDF RGB channels. Bottom row shows renderings of reference BRDF (left) and reconstructions for the respective icosphere levels.

under various error influences are reported in table 2 using the ΔE_{00} error measure.

4.2. Influence of Vertex Position Noise

Errors in triangulation during the SL scanning procedure directly affect the precision of vertex positions. Commonly, but depending on material, SL scanners have a very high precision in the order of microns [4]. To investigate the sensitivity to vertex positions, all vertices are affected by three relatively large levels of noise prior to projection onto HDR images. The noise is added as a normally distributed noise on the xyz-components of each vertex with standard deviations of $\sigma \in \{1, 3, 5\}$mm. In figure 5, the qualitative evaluations for material "blue-rubber" are presented. As is apparent, grazing angle behavior is greatly affected by vertex uncertainties. This is most likely caused by the fact that even small uncertainties may at grazing angles project a vertex onto the black background, rather than the target sample. Likewise, for very specular materials as "specular-black-phenolic", vertices may miss the very narrow highlight causing errors in estimating the specular reflection.

		blue-rubber	green-metallic-paint	purple-paint	specular-black-phenolic
Ideal		0.41 ± 0.67	3.00 ± 3.23	0.55 ± 0.75	1.43 ± 1.92
Vertex	1mm	0.66 ± 1.15	3.03 ± 3.44	0.91 ± 1.56	2.27 ± 3.37
	3mm	2.50 ± 2.82	3.11 ± 4.17	2.11 ± 3.16	3.41 ± 5.15
	5mm	4.16 ± 4.08	3.24 ± 4.77	3.59 ± 4.32	4.22 ± 5.76
Normal	5°	0.42 ± 0.63	3.00 ± 3.20	0.68 ± 1.03	3.15 ± 5.66
	10°	0.67 ± 0.98	3.08 ± 3.17	0.99 ± 1.70	3.72 ± 6.92
	30°	1.81 ± 2.11	4.70 ± 5.08	2.27 ± 3.27	5.24 ± 8.31
Light Pos.	10mm	0.51 ± 0.72	3.01 ± 3.26	0.62 ± 0.84	2.17 ± 3.32
	25mm	0.73 ± 0.89	2.68 ± 3.03	1.02 ± 1.16	3.05 ± 5.03
	50mm	1.77 ± 1.92	3.14 ± 3.51	1.91 ± 2.27	3.74 ± 6.08
Light Int.	5%	0.64 ± 0.79	3.01 ± 3.29	0.66 ± 0.78	1.90 ± 2.80
	10%	1.00 ± 1.15	3.05 ± 3.38	0.96 ± 0.99	1.98 ± 2.81
	20%	1.75 ± 1.75	3.40 ± 3.74	1.86 ± 2.10	2.27 ± 2.80

Table 2: Errors for different types of noise introduced to the structured light scanner system. Errors are measured as the average ΔE_{00} color difference between tone mapped renderings of ground truth BRDF and reconstruction.

4.3. Influence of Vertex Normal Noise

As surface normals are often derived from the mesh, they often suffer from large uncertainty. This directly affects the frame of reference in which the BRDF is estimated. To simulate such uncertainties, all normals in the mesh are tilted in a random direction away from the true normal by a normally distributed angle. Three different standard deviations are reported here: $\sigma \in \{5^\circ, 10^\circ, 30^\circ\}$. In figure 6, qualitative evaluations are presented for "purple-paint". Although specular highlights are somewhat affected, it is noteworthy how large an amount of noise we can add to the normals while still obtaining a decent recovery of the material.

4.4. Influence of Light Source Position Noise

As mentioned in section 3.2.2, it may be difficult to determine the precise position of light sources in the SL system. To simulate such uncertainties, normally distributed noise is added to the xyz-components of the light positions (projector and scene-light) with standard deviations of $\sigma \in \{10, 25, 50\}$ mm. In figure 7, the influence of this error is shown for the "green-metallic-paint" material. Surprisingly, even for the relatively large amounts of noise applied here, reconstructions remain very close to the results under ideal conditions as well as the ground truth.

4.5. Influence of Light Source Intensity Noise

Finally, noise applied to the intensity of the light sources (projector and scene light) is applied. Here, the noise is modeled as a normally distributed percentage with a mean of 100%. The standard deviation of the noises applied are $\sigma \in \{5, 10, 20\}\%$. Figure 8 shows the results for the material "specular-black-phenolic". Here, the strong prior of the BRDF reconstruction model almost fully handles the uncertainties in intensity although this property is very tightly coupled to reflectance.

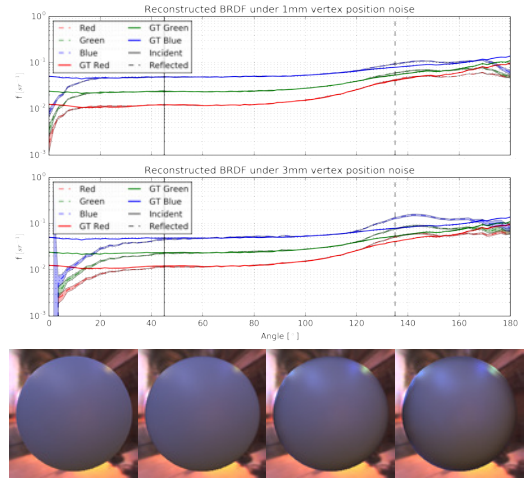


Figure 5: BRDF reconstructions of "blue-rubber" material, for increasing noise added to vertex positions, shown as 45° in-plane profiles. We add Gaussian noise with a standard deviation of $\{1, 3, 5\}$mm. BRDF RGB channels are plotted with solid lines as ground truth and dashed lines as the mean reconstruction. Shaded regions indicate limits for ± 2 standard deviations. Bottom row shows renderings of reference BRDF (left) and mean reconstructions for the respective noise levels. Statistics are based on 30 evaluations.

4.6. Summary

Table 2 summarizes the BRDF errors caused by introducing the noise types listed above using the ΔE_{00} error measure. We observe that, not surprisingly, accuracy of vertex positions has a great impact on the quality of the recovered material. Recall that the object size is 100 mm, only a few percent error are enough to throw the BRDF estimate off. On the contrary, variations in surface normals are less influencing than we would have expected, requiring especially for soft materials a lot of noise before throwing the BRDF recovery off. Finally positions and intensities of light sources are seen to have a surprisingly small impact on BRDF reconstructions.

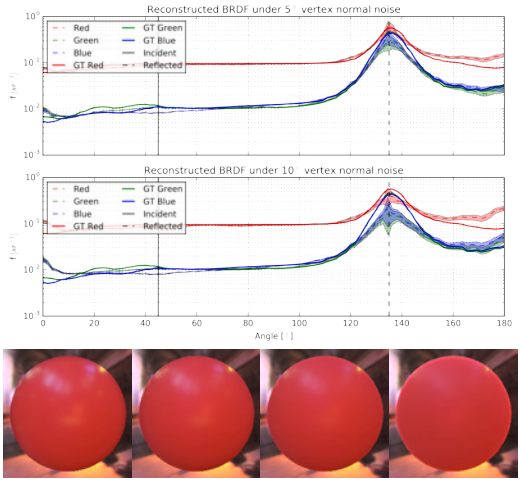


Figure 6: BRDF reconstructions of "purple-paint" material, for increasing noise added to vertex normals, shown as $45°$ in-plane profiles. We add Gaussian noise with a standard deviation of $\{5°, 10°, 30°\}$.

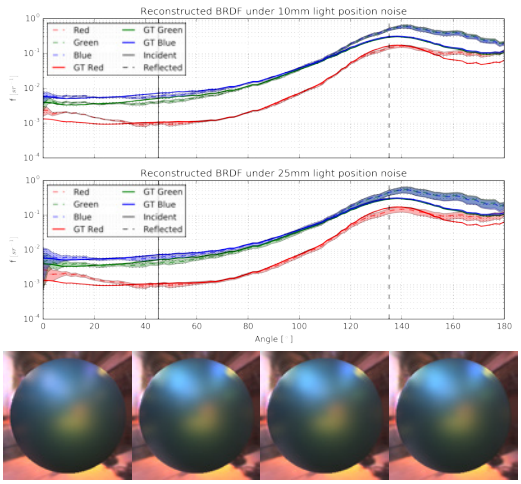


Figure 7: BRDF reconstructions of "green-metallic-paint" material, for increasing noise added to the two light source positions, shown as $45°$ in-plane profiles. We add Gaussian noise with a standard deviation of $\{10, 25, 50\}$mm.

5. Discussion and Conclusion

We investigated how a structured light 3D scanning system can be modified with minimal effort to also estimate

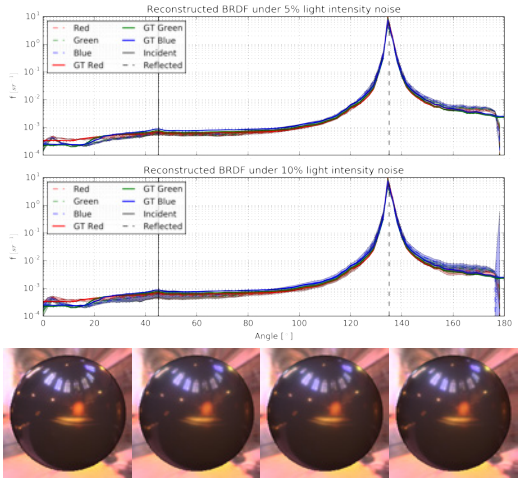


Figure 8: BRDF reconstructions of "specular-black-phenolic" material, for increasing noise added to the two light source intensities, shown as $45°$ in-plane profiles. We add Gaussian noise with a standard deviation of $\{5\%, 10\%, 20\%\}$.

BRDFs. Results indicate that high quality reflectance recovery is in fact possible in such a setup. We carried out a variational study in a simulated environment to investigate how a range of uncertainties in system parameters affect the quality of the estimated reflectance properties. The goal of this study is to provide system designers with a lookup table of system parameter uncertainties required to recover a given material at a given quality-level. This is needed in the design phase of future systems for full appearance acquisition. Tables 1 and 2 provide this information and demonstrate that even under the poor gonioreflectometric conditions provided by a SL system, very high quality reflectance may be recovered. An interesting insight gained here is that uncertainties in surface normals in fact have a smaller impact on the quality of estimated BRDFs than one might have expected. Likewise, uncertainties in illumination properties, including position and intensity, have little influence on the recovered reflectance.

Although the experiments carried out here are only simulated, we believe that they reflect well what can be expected from real world measurements. It has not been the intention with this paper to cover the physical implementation of this pipeline as well as the performance of the approach in real-world scenarios. Nonetheless, the images presented in figure 2 do in fact originate from an actual implementation of the system, demonstrating that it also works in practice. It is our intention to elaborate on these results in the future.

References

[1] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers. *Advanced High Dynamic Range Imaging: Theory and Practice*. AK Peters (CRC Press), Natick, MA, USA, 2011.

[2] J. F. Blinn. Models of light reflections for computer synthesized pictures. *Proceedings of ACM SIGGRAPH 77*, 11(2):192–198, July 1977.

[3] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of SIGGRAPH 97*, pages 369–378. ACM/Addison-Wesley, 1997.

[4] E. R. Eiríksson, J. Wilm, D. B. Pedersen, and H. Aanæs. Precision and accuracy parameters in structured light 3-D scanning. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5/W8:7–15, 2016.

[5] J. Geng. Structured-light 3d surface imaging: a tutorial. *Advances in Optics and Photonics*, 3(2):128–160, 2011.

[6] L. Gomes, O. R. P. Bellon, and L. Silva. 3D reconstruction methods for digital preservation of cultural heritage: A survey. *Pattern Recognition Letters*, 50:3–14, December 2014.

[7] M. Halioua and H.-C. Liu. Optical three-dimensional sensing by phase measuring profilometry. *Optics and Lasers in Engineering*, 11(3):185–215, 1989.

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[9] S. Hill, S. McAuley, A. Conty, M. Drobot, E. Heitz, C. Hery, C. Kulla, J. Lanz, J. Ling, N. Walster, F. Xie, A. Micciulla, and R. Villemin. Physically based shading in theory and practice. In *ACM SIGGRAPH 2017 Courses*, July 2017.

[10] M. Holroyd, J. Lawrence, and T. Zickler. A coaxial optical scanner for synchronous acquisition of 3d geometry and surface reflectance. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2010)*, 29(4):99:1–99:12, July 2010.

[11] K. Ikeuchi and K. Sato. Determining reflectance properties of an object using range and brightness images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(11):1139–1153, November 1991.

[12] J. Krzesłowski, R. Sitnik, and G. Mączkowski. Integrated three-dimensional shape and reflection properties measurement system. *Applied optics*, 50(4):532–541, February 2011.

[13] E. P. Lafortune, S.-C. Foo, K. E. Torrance, and D. P. Greenberg. Non-linear approximation of reflectance functions. In *Proceedings of SIGGRAPH 1997*, pages 117–126. ACM/Addison-Wesley, 1997.

[14] H. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics*, 22(2):234–257, April 2003.

[15] J. Lu and J. Little. Reflectance function estimation and shape recovery from image sequence of a rotating object. In *Proceedings of International Conference on Computer Vision (ICCV 1995)*, pages 80–86. IEEE, 1995.

[16] S. R. Marschner, S. H. Westin, E. P. Lafortune, K. E. Torrance, and D. P. Greenberg. Image-based BRDF measurement including human skin. In *Rendering Techniques 1999*, pages 131–144. Springer, 1999.

[17] W. Matusik, H. Pfister, M. Brand, and L. McMillan. A data-driven reflectance model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2003)*, 22(3):759–769, 2003.

[18] J. B. Nielsen, E. R. Eiríksson, R. L. Kristensen, J. Wilm, J. R. Frisvad, K. Conradsen, and H. Aanæs. Quality assurance based on descriptive and parsimonious appearance models. In *Workshop on Material Appearance Modeling (MAM 2015)*, pages 21–24. The Eurographics Association, June 2015.

[19] J. B. Nielsen, J. R. Frisvad, K. Conradsen, and H. Aanæs. Addressing grazing angle reflections in Phong models. In *SIGGRAPH Asia 2014 Posters*, page 43. ACM, 2014.

[20] J. B. Nielsen, H. W. Jensen, and R. Ramamoorthi. On optimal, minimal BRDF sampling for reflectance acquisition. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2015)*, 34(6):186:1–186:11, November 2015.

[21] M. Pharr, W. Jakob, and G. Humphreys. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann, third edition, 2016.

[22] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, June 1975.

[23] S. Rusinkiewicz. A new change of variables for efficient BRDF representation. In *Rendering Techniques '98 (Proceedings of EGWR 1998)*, pages 11–22, 1998.

[24] Y. Sato, M. D. Wheeler, and K. Ikeuchi. Object shape and reflectance modeling from observation. In *Proceedings of SIGGRAPH 1997*, pages 379–387. ACM/Addison-Wesley, 1997.

[25] C. Schwartz, R. Sarlette, M. Weinmann, and R. Klein. DOME II: A parallelized BTF acquisition system. In *Workshop on Material Appearance Modeling (MAM 2013)*, pages 25–31. The Eurographics Association, June 2013.

[26] G. Sharma, W. Wu, and E. N. Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

[27] R. Sitnik, J. KrzesŁ, M. Grzegorz, et al. Archiving shape and appearance of cultural heritage objects using structured light projection and multispectral imaging. *Optical Engineering*, 51(2):021115–1, 2012.

[28] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America*, 57(9):1105–1114, September 1967.

[29] B. Tunwattanapong, G. Fyffe, P. Graham, J. Busch, X. Yu, A. Ghosh, and P. Debevec. Acquiring reflectance and shape from continuous spherical harmonic illumination. *ACM Transactions on graphics (Proceedings of SIGGRAPH 2013)*, 32(4):109, 2013.

[30] M. Weinmann and R. Klein. Advances in geometry and reflectance acquisition (course notes). In *Proceedings of SIGGRAPH Asia 2015 Courses*. ACM, November 2015.

[31] L. Williams. Casting curved shadows on curved surfaces. *Computer Graphics (Proceedings of SIGGRAPH 78)*, 12(3):270–274, August 1978.

[32] Z. Xu, J. B. Nielsen, J. Yu, H. W. Jensen, and R. Ramamoorthi. Minimal BRDF sampling for two-shot near-field reflectance acquisition. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016)*, 35(6):188:1–188:12, November 2016.

[33] S. Zhang and P. S. Huang. Novel method for structured light system calibration. *Optical Engineering*, 45(8):083601–083601, 2006.

[34] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

Contribution C

Wearable Gaze Trackers: Mapping Visual Attention in 3D

Wearable Gaze Trackers: Mapping Visual Attention in 3D

Rasmus R. Jensen[1], Jonathan D. Stets[1( )], Seidi Suurmets[2], Jesper Clement[2], and Henrik Aanæs[1]

[1] Technical University of Denmark, Kongens Lyngby, Denmark
stet@dtu.dk
[2] Copenhagen Business School, Frederiksberg, Denmark

Abstract. The study of visual attention in humans relates to a wide range of areas such as: psychology, cognition, usability, and marketing. These studies have been limited to fixed setups with respondents sitting in front of a monitor mounted with a gaze tracking device. The introduction of wearable mobile gaze trackers allows respondents to move freely in any real world 3D environment, removing the previous restrictions.

In this paper we propose a novel approach for processing visual attention of respondents using mobile wearable gaze trackers in a 3D environment. The pipeline consists of 3 steps: modeling the 3D area-of-interest, positioning the gaze tracker in 3D space, and 3D mapping of visual attention.

The approach is general, but as a case study we created 3D heat maps of respondents visiting supermarket shelves as well as finding their in-store movement relative to these shelves. The method allows for analysis across multiple respondents and to distinguish between phases of in-store orientation (far away) and product recognition/selection (up close) based on distance to shelves.

1 Introduction

The study of human visual attention relates to a wide range of areas such as: psychology, cognition, useability, and marketing. In order to directly study this in various settings, eye tracking has become a standard method. A common way of visualizing and analysing gaze data is using Areas Of Interest (AOI) and attentional heat maps [13]. The heat maps represent the spatial distribution of eye movement throughout the AOI and can often be used for quantitative analysis. The most common method of visualizing heat maps is using a Gaussian based solution. Here, four parameters are used to determine the appearance of the heat map: the width of the basic construct, the use of fixations vs. raw data, whether accounting for duration of fixation and the mapping color altitude form [3]. For many years, mapping visual attention as heat maps has been limited to static setups with respondents sitting in front of a screen mounted with a stationary calibrated gaze tracker. Such a setup can accurately map the visual attention as a heat map of what is projected on the screen, but obviously limits

P. Sharma and F.M. Bianchi (Eds.): SCIA 2017, Part I, LNCS 10269, pp. 66–76, 2017.
DOI: 10.1007/978-3-319-59126-1_6

Fig. 1. Supermarket vegetables shown as a 3D model with heatmap and respondent viewing points.

the visual attention to a 2D surface. The recent introduction of mobile wearable gaze trackers (Fig. 2) enables data collection in about any real-world environment. On mobile wearable eye-trackers, the scene is recorded using a front facing camera, and gaze data collected from eye tracking cameras can be projected onto this video. Despite the potential of introducing recordings of three dimensional scenes, common for both the stationary and mobile wearable eye-tracker is that ultimately the data is still recorded and analysed in 2D.

Mapping visual attention data recorded in a 3D space to a 2D heatmap is not straightforward. A simple approach is to find the best homographic correspondence between a reference image and a given frame from the eye-tracker, and then map the gaze according to this homography [4,12]. Figure 3 shows common errors in mapping using a homography relative to the actual mapping onto a 3D AOI. We argue that gaze collected in 3D mapped onto a 2D reference image using a homography will always be limited as a result of incorrect mappings.

Fig. 2. Tobii Pro Glasses 2 [12]. A wearable gaze tracker that tracks a respondents eye movements using IR cameras, while also recording the environment with a front facing video camera.

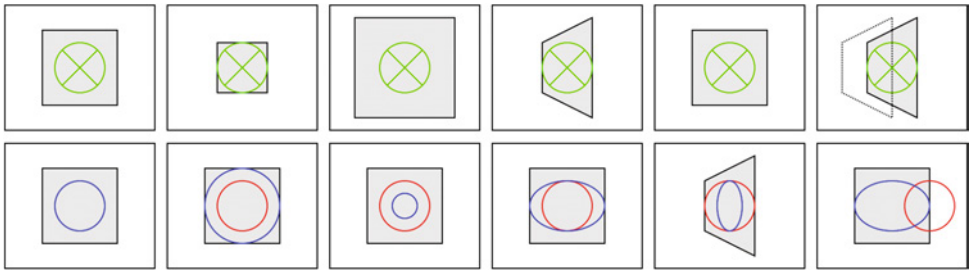


Fig. 3. This figure shows the common errors relating to mapping gaze. The top row shows a respondents viewpoint of an AOI with the gaze point in green, while the second row shows a reference view of an AOI with gaze mapped as homography mapping in red and mapping according to 3D structure in blue. First column shows mapping, when the viewpoint of the respondent and reference coincide. In this case, homography mapping and mapping according to 3D structure will be identical and perfectly overlapping. Column 2 and 3 show the mapping, when respondent is closer to or further from the AOI. Using homography mapping, the gaze point does not change along with the movement of the respondent. Row 4 and 5 show how the mapped gaze changes shape according to changes in viewpoint for the 3D mapping, while homography mapping does not change accordingly. The final column shows the error, when the homography is offset from the plane of the actual viewpoint, which introduces parallax error. (Color figure online)

We propose a solution to these problems and limitations by modelling an AOI in 3D as a reference for mapping gaze data. The reference model is reconstructed from photographs of the AOI to establish a good base for image feature matching and a high quality model mesh. We demonstrate a fully automatic pipeline for generating a 3D attention heat map, and furthermore the possibility of calculating the respondent viewing points as shown in Fig. 1. Our pipeline enables spatial filtering, positioning and orientation relative to the selected AOI, as well as correlation of multi-respondent data. We use supermarket shelves as a case study, but our pipeline is not limited to this setup. Our method requires a standard digital camera to capture images of the reference model, and a wearable gaze tracker with a front facing camera, such as the one shown in Fig. 2, for recording the scene and gaze data.

There are a number of recent studies that addresses the need to move mapping of visual attention to 3D. [11] introduces the potential of measuring 3D gaze coordinates from head-mounted gaze trackers, and [9] proposes visualisation of 3D gaze data on to virtual computer generated models. A method similar to our pipeline is described in [10], which demonstrates the use of a Microsoft Kinect to create a 3D reference model. Our method differs by using images to create a more dense point cloud, which also enables us to backproject the heat map to a traditional 2D visualization for comparison.

2 Data

We have collected data in both a real world supermarket and using a mock-up supermarket shelf in our lab. Reference data of the AOIs have been captured using a digital mirrorless camera: a Panasonic GH4 with a 12 mm lens (24 mm in 35 mm equivalent). To collect respondent data we have used the Tobii Pro Glasses 2 wearable gaze tracker [12] (Fig. 2), which collects the respondents view using a front facing video camera, while also recording the respondent gaze direction using 4 infrared cameras facing the eyes. Both cameras were calibrated using a standard checkerboard approach [16]. Data was collected of four in-store product sections in a supermarket: wine, vegetables, flour and cereal, as well as a mock-up of the cereal section in our lab. We used the digital camera to capture sets of reference images to cover the desired AOIs (12–20 images of each AOI). Gaze and video data were collected of respondents visiting the given sections (16 sets), visiting the store but acquired to get cereal (4 sets), and finally, presented for a mock-up of the cereal section in the lab (6 sets). All gaze data samples are raw, so no fixation filtering has been applied [3].

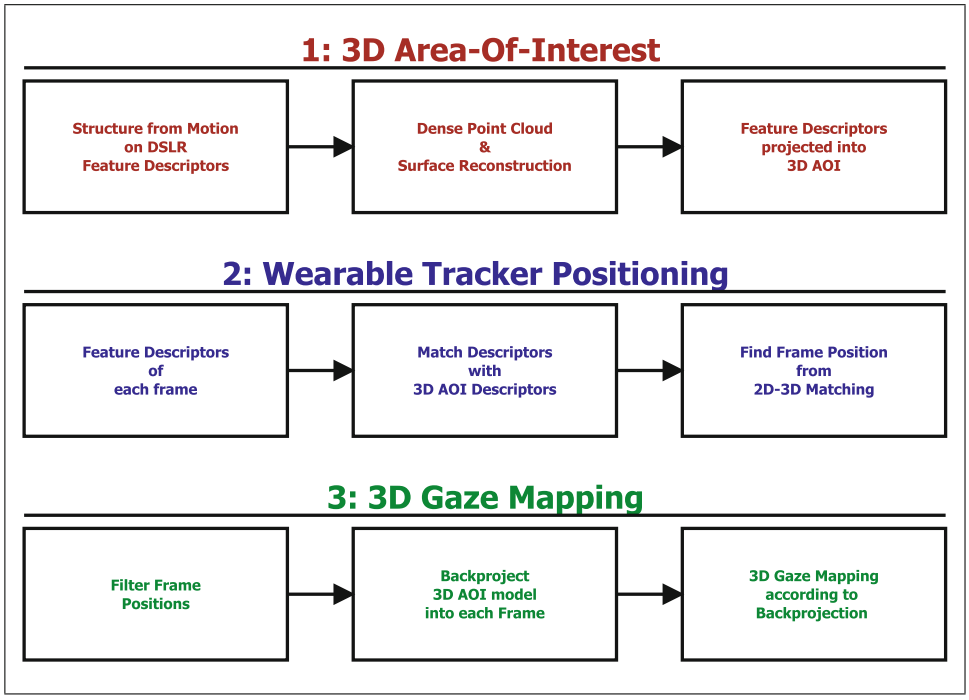


Fig. 4. The 3 steps in our proposed pipeline to construct 3D gaze mapping: Modelling of an Area-Of-Interest, Eye-tracker frame positioning, and finally the gaze mapping.

3 Method

In order to map gaze data onto a 3D AOI, we propose a pipeline consisting of three parts (Fig. 4): construction of the 3D AOI reference model, localization of the wearable gaze tracker frames relative to the reference model, and finally gaze mapping onto the AOI as a heat map.

3.1 Modelling a 3D Area-Of-Interest

The 3D AOI reference model is built using a series of images of the AOI. This task is divided further into three steps (Fig. 4). First, we use structure from motion to find the spatial camera positions and a sparse point cloud representation. We have opted for a structure from motion (SfM) [6] implementation, which requires a sequence of images followed by an image rectification based on the parameters obtained from the camera calibration. SIFT descriptors [7,15] are found in each image and sequentially matched across the sequence of images in an iterative fashion. Images with sufficient feature matches are included, while the extrinsic camera parameters are estimated and refined using bundle adjustment [14].

Given the estimated extrinsic camera parameters, we move onto dense point cloud estimation using the patch-expansion approach to multiview stereopsis proposed by Furukawa and Ponce [2]. This method robustly produces dense representations from which a surfaces are reconstructed using Poisson surface reconstruction by Kazhdan et al. [5]. A 3D modelled AOI from the cereal section in a supermarket is shown in Fig. 5(a).

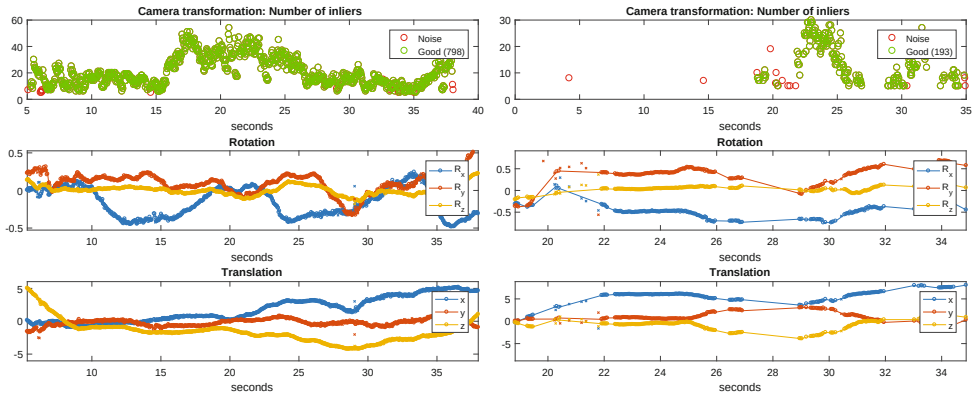
As a preparation step for the localization of the wearable gaze tracker later in the pipeline, we use backprojection with depth management of the 3D AOI to project the model into each reference image. This is done in order to project 2D SIFT descriptors [7] into the 3D space, allowing the 2D descriptors between each frame from the gaze tracker and 3D AOI to be compared.

(a) 3D Area-Of-Interrest reference model. (b) 3D AOI backprojected onto a gaze tracker frame.

Fig. 5. The 3D AOI in (a) is backprojected onto an undistorted gaze tracker frame and the gaze point with trace from previous frames (b). The frame is shown in black and white, while the projection is shown in color.

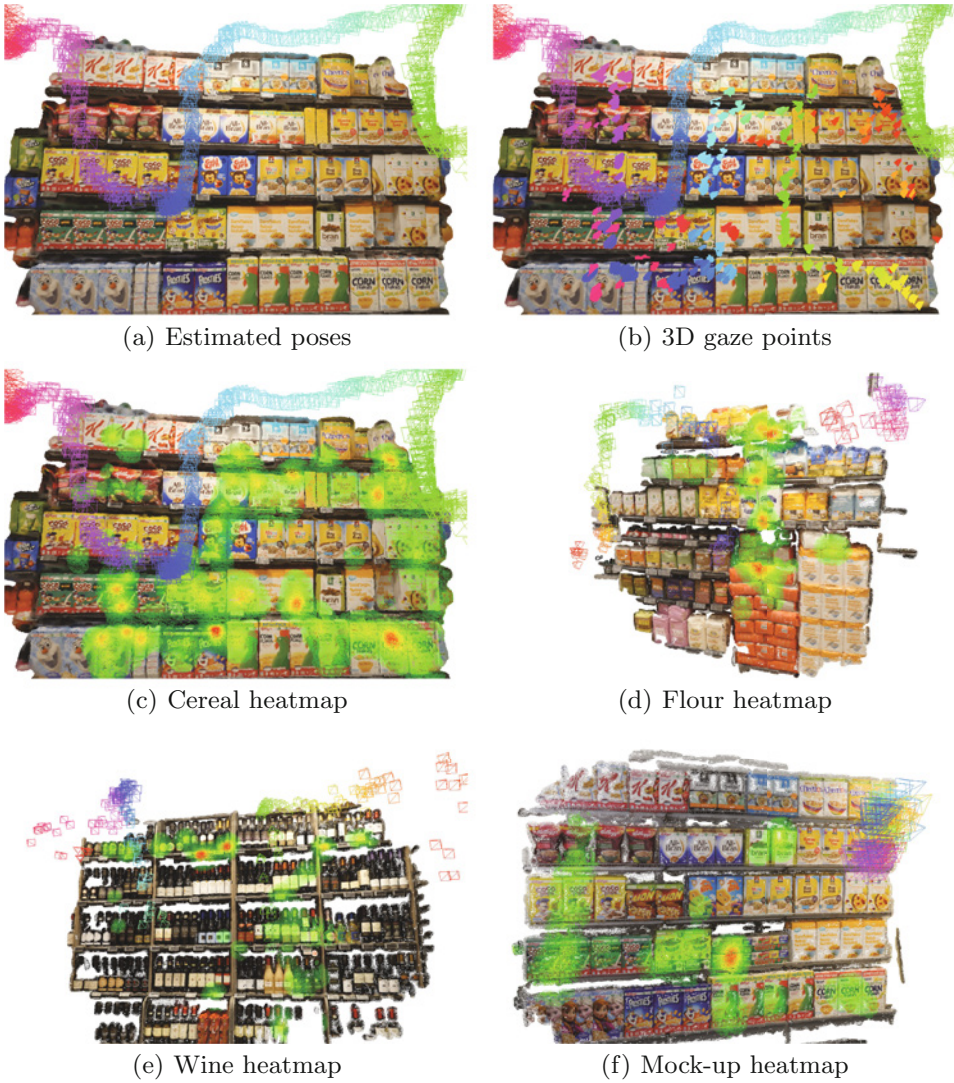
3.2 Wearable Gaze Tracker Frame Localization

In order to correctly map gaze data on the 3D AOI, each frame from the gaze tracker has to be positioned relative to the AOI (if visible). The SIFT descriptors in each frame are matched with the reference 2D descriptors projected into 3D space, when constructing the 3D AOI. These correspondences are sent to a 2D to 3D pose solver, which finds the best fit camera pose using a RANSAC approach discarding outliers [1]. Given sufficient corresponding points, the solver will return the correct camera pose relative to the AOI. Without sufficient good matches the solver either fails or returns a false camera pose. Since a given frame might not cover any part of the AOI, the resulting matching consists of either a lot of true positive or a few false positive matches. Figure 5(b) shows the 3D AOI backprojected into a frame from the wearable tracker using the estimated camera pose. This backprojection is an immediate sanity check, showing the correctness of the pose estimation. Incorrect pose estimates tends to be very inconsistent from one frame to the next. To speed things up we have used the above approach to find the pose in keyframes, which are followed by frames, where the correspondence points are tracked using optical flow [8] (1 keyframe followed by 5 optical flow frames). This is substantially faster than finding and matching features in each frame. The pose solver is initialized with the pose from the previous frame, which along with the optical flow gives timewise consistency in the pose estimation.



(a) Cereal shelves (b) Flour shelves

Fig. 6. Estimated respondent poses from visits to the cereal and flour shelves with timestamp. First row of plot: the framewise number of inliers (y-axis) in the camera positioning solver. Green points are included as reliable and red points are considered noise. Noise points are filtered out based on spatial and rotational inconsitency. Second and third row of plots are the rotation and translation with inliers shown as a connected graph and the few outliers as single points. (Color figure online)



(a) Estimated poses (b) 3D gaze points

(c) Cereal heatmap (d) Flour heatmap

(e) Wine heatmap (f) Mock-up heatmap

Fig. 7. Poses, gaze points and heatmaps obtained from the data of the 5 sections included in our study.

3.3 Mapping Gaze Data

The pose estimation is unbiased and may results in a few faulty poses. We consider these noise and use the following approach to filter them from the good poses. Correct pose estimates between consecutive frames is assumed to have small variation, while incorrect poses are very inconsistent. This inconsistency is used to identify and discard faulty poses. In Fig. 6 the good pose estimates is shown as a connected graphs with discarded poses as outlying points. The number of inliers returned from the 2D to 3D pose solver is a good estimator of correctness, but thresholding this number is not as robust as filtering the pose.

A respondent moving in front of the AOI is shown in Fig. 7(a). Using the good poses, the gaze can be mapped onto the 3D AOI model creating a 3D heat map as seen in Fig. 7(b). The gaze intersection with the 3D model is found using backprojection with depth management into the current frame, which is significantly faster than calculating the intersection between the line of sight and the 3D model.

A similar approach is taken, when creating the heatmap. Here a predetermined symmetric 2D Gaussian function with center at the gaze coordinate is added to a sum map of Gaussians in 3D space. Using a Gaussian serves both the fact that sight is not an infinitely small point, while also incorporating some uncertainty in the gaze estimates. Discussions about the size of the Gaussian, and whether the raw gaze data or fixation filtered data should be used is beyond the scope of this work. The resulting heatmaps visualized on the 3D AOIs can be seen in Figs. 1 and 7(c) to (e).

One benefit worth noticing is, that the approach of mapping Gaussian to the backprojection of the AOI allows for a normalization of the contribution from each gaze point. It also addresses the problems shown in Fig. 3. When a respondent is close, the covering of the Gaussian gaze point of the 3D AOI will be small with a locally high intensity. Respondents far away will cover a larger area in the 3D AOI, which will result in less locally intensive mapping. It also handles change in perspective, while effectively shaping the Gaussian according to the viewpoint without introducing parallax error. Since the sum of Gaussian gaze points is done on a 3D model, the heatmap can be projected into any frame or reference image. The backprojection of the a heatmap is shown in Fig. 8(a) as an overlay to the original image.

(a) 3D heatmap backprojected into reference image

(b) 2D heatmap from iMotions software

Fig. 8. Heatmaps based on 3D gaze mapping and 2D gaze mapping. For the 3D mapping the heatmap has been backprojected into the DSLR frame used for the 2D mapping.

4 Results

The core of our presented pipeline is the ability to correctly find the pose of the wearable gaze tracker relative to the 3D AOI in a given frame. Validating this after filtering puts each frame in one of four categories:

True positive correctly detecting the AOI.
True negative correctly not detecting the AOI.
False positive incorrect detection of the AOI.
False negative incorrectly not detecting the AOI.

Reviewing the output videos with 3D AOI overlay backprojected as presented in Fig. 5(b) is an easy way to quickly asses the quality of the AOI detection. Such a review shows non or only a very few false positives, but some false negatives. Since the gaze tracker has a very small sensor, the sensor struggles with low indoor light, which results in both frames with motion blur from head movement and rolling shutter. In the supermarket setting, these frames provide the vast majority of false negatives and one could debate, whether they are actually false negatives. Occlusion from people or other shelves can also cause false negatives. Reviewing both the frame positions as a graph in Fig. 6 or the resulting spatial positions in Fig. 7(a) are can also provide quick qualitative verifications in addition to reviewing a video with backprojected 3D AOI.

We have reprojected the heatmap into a reference image, which has also been applied homography gaze mapping using the iMotions 6.2 software [4] and the results are shown in Fig. 8. The heatmaps are both based on raw data samples but using different techniques: 3D mapping and homography mapping respectively. This means they cannot be compared directly, however there have clear similarities of the path pattern and duration of attention.

5 Conclusion

We have successfully created 3D AOIs and heat maps for respondents visiting the five sections in our data set: vegetables, cereal, flour, wine, and cereal mock-up. Our proposed pipeline does away with the problems relating to mapping gaze using a homography. The proposed pipeline is fully automatic and runs at $\sim$2 fps using a combination of Matlab, mex, vlfeat and OpenCV. A full C++ implementation will provide further speed up, but the scope was to demonstrate a feasible pipeline, which allows researchers to spend time only on the results, once processed. Concerns such as visual attention mapped based on fixation vs. raw data, size of the Gaussian gaze point, and cross respondent analysis has not been evaluated. We found that our pipeline works well in in-store settings, since store products tend to have very distinct image features. However, settings with only repeating image features, such as frames with only the same product present, complicates the feature matching. This is often the case, when the respondent is very near a product shelf. On the other hand, detection works well in the case where the respondent is inspecting the shelves at an arm length

distance, which in many cases is the important frames for generating heat maps. Our approach provides a fully automatic method of mapping gaze data and positioning of the respondent relative to the AOI, thus adding another dimension to the resulting data.

Acknowledgements. This work has been funded by the Innovation Fund (Denmark) and carried out in collaboration with iMotions. We would like to thank both for the support and collaboration.

References

1. Bradski, G.: The OpenCV Library. Dr. Dobb's J. Softw. Tools, Article id 2236121 (2000). http://code.opencv.org/projects/opencv/wiki/CiteOpenCV
2. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. **1**(1), 1–14 (2007). ISSN 01628828. http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.161
3. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye Tracking: A Comprehensive Guide to Methods and Measures. OUP, Oxford (2011)
4. iMotions. iMotions biometric research platform (2017). https://imotions.com/
5. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Son Geometry Processing (SGP 2006), pp. 61–70, Aire-la-Ville, Switzerland. Eurographics Association (2006). ISBN 3-905673-36-3
6. Koenderink, J., Van Doorn, A.: Affine structure from motion. Optical Soc. Am. A **8**(2), 337–385 (1991). doi:10.1364/JOSAA.8.000377, ISSN 1084–7529
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
8. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision (1981). ISSN 17486815, http://www.ic.unicamp.br/rocha/teaching/2013s1/mc851/aulas/additional-material-lucas-kanade-tracker.pdf
9. Maurus, M., Hammer, J.H., Beyerer, J.: Realistic heatmap visualization for interactive analysis of 3d gaze data. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 295–298. ACM (2014)
10. Paletta, L., Santner, K., Fritz, G., Mayer, H., Schrammel, J.: 3d attention: measurement of visual saliency using eye tracking glasses. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 199–204. ACM (2013)
11. Pfeiffer, T.: Measuring and visualizing attention in space with 3d attention volumes. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 29–36. ACM (2012)
12. Tobii Pro. Tobii pro glasses 2 (2017). http://www.tobiipro.com/product-listing/tobii-pro-glasses-2/
13. Purucker, C., Landwehr, J.R., Sprott, D.E., Herrmann, A.: Clustered insights: improving eye tracking data analysis using scan statistics. In: Psychological Considerations on Car Designs-An Investigation of Behavioral and Perceptual Aspects Using Eye Tracking and Cross-Cultural Studies (2012)
14. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment — a modern synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) IWVA 1999. LNCS, vol. 1883, pp. 298–372. Springer, Heidelberg (2000). doi:10.1007/3-540-44480-7_21, http://www.springerlink.com/content/plvcrq5bx753a2tn

15. Vedaldi, A., Fulkerson, B.: Vlfeat. In: Proceedings of the International Conference on Multimedia (MM 2010), vol. 3, no. 1, p. 1469 (2010). doi:10.1145/1873951.1874249, http://dl.acm.org/citation.cfm?doid=1873951.1874249
16. Zhang, Z.: Flexible camera calibration by viewing a plane from unknown orientations. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 1, no. c, p. 7 (1999). doi:10.1109/ICCV.1999.791289, ISSN 01628828

Contribution D

Visualization and labeling of point clouds in virtual reality

Visualization and Labeling of Point Clouds in Virtual Reality

Jonathan Dyssel Stets
Technical University of Denmark

Yongbin Sun
Massachusetts Institute of Technology

Wiley Corning
MIT Media Lab

Scott W. Greenwald
MIT Media Lab

Figure 1: Screen capture from the Virtual Reality application. A point cloud of jugglers is displayed, the left and right images show the point cloud before and after the juggling pins have been annotated by the user.

ACM Reference Format:
Jonathan Dyssel Stets, Yongbin Sun, Wiley Corning, and Scott W. Greenwald. 2017. Visualization and Labeling of Point Clouds in Virtual Reality. In *Proceedings of SA '17 Posters, Bangkok, Thailand, November 27-30, 2017,* 2 pages.
https://doi.org/10.1145/3145690.3145729

1 INTRODUCTION

We present a Virtual Reality (VR) application for labeling and handling point cloud data sets. A series of room-scale point clouds are recorded as a video sequence using a Microsoft Kinect. The data can be played and paused, and frames can be skipped just like in a video player. The user can walk around and inspect the data while it is playing or paused. Using the tracked hand-held controller, the user can select and label individual parts of the point cloud. The points are highlighted with a color when they are labeled. With a tracking algorithm, the labeled points can be tracked from frame to frame to ease the labeling process. Our sample data is an RGB point cloud recording of two people juggling with pins. Here, the user can select and label, for example, the juggler pins as shown in Figure 1. Each juggler pin is labeled with various colors to indicate different labels.

SA '17 Posters, November 27-30, 2017, Bangkok, Thailand

ACM ISBN 978-1-4503-5405-9/17/11.
https://doi.org/10.1145/3145690.3145729

2 MOTIVATION

We consider the use case of viewing animated point cloud data in VR. Although individual frames appear grainy and low-resolution, when they are animated, they produce a lifelike effect that is suitable for representing dynamic real-world scenes. However, without any labeling or metadata, it is difficult to edit or modify these scenes. Labeling a person with a name, or changing the color of an object would be examples of basic editing tasks. Segmenting the point clouds in each frame, and establishing how the segments in each frame correspond with those in surrounding frames, makes it possible to perform such tasks.

Furthermore, the same set of user interface affordances can be used for other use cases as well. The demand for large labeled point cloud data sets is increasing as the need for classification algorithms for such data is gaining popularity, especially for data-driven approaches like deep neural networks. Navigation of self driving cars, social robots and robot interaction in general often rely on 3D data that needs to be classified or segmented. Additionally, the use of 3D sensors is increasing, while these sensors are getting more mobile, better and more available. This enables the potential for more and better 3D point cloud data. The website Semantic3D [Semantic3d 2016] provides examples of large labeled 3D point cloud data sets. Oftentimes, such 3D data sets are labeled manually using a 2D computer monitor and a computer mouse or similar input device. We argue that a two-dimensional human-computer interaction with three-dimensional data can be inefficient in many cases. For example, the labeling problem is more difficult with a 2D interface when the orthogonal *xy*, *xz*, or *yz* planes do not separate the points to be labeled from those surrounding them.

Figure 2: The juggling pin is being labeled with a blue label using the hand-held controller. Haptic feedback in the controller is provided when points are selected.

We present an application that demonstrates an intuitive way to inspect and label 3D point cloud data. The user simply inspects the data by moving around in the 3D point cloud, as one would do in the real world. Labeling is done by pointing and drawing in the 3D space with a hand-held controller to mark relevant parts of the data.

3 DETAILS

We have recorded a data set using a Microsoft Kinect 2 and a custom-built software tool. The software records and converts the depth map and rgb video stream from the Kinect to a point cloud and stores the (X, Y, Z, R, G, B) data in a binary format. The point clouds are recorded and stored approximately 30 times per second. The set of point clouds is loaded into the VR application and visualized at room scale. We use the HTC Vive setup for VR, which enables tracking of both headset and the hand-held controllers. Using the touchpad on the hand-held controller, the user can navigate back and forward in the point cloud frames. A marker-tool is attached to the controller and this can be used like a paint brush to label points in the point cloud (see Figure 2). Haptic feedback is provided when selecting the points to add the feeling of physically interacting with the data. Individual parts of the point cloud can be labeled with different labels, each represented by varying colors. The user can switch between labels on the controller touchpad, as well as choosing an eraser-tool to unlabel points in the point cloud. We argue that interacting with 3-dimensional data in this way is an easy and fast alternative to a traditional 2-dimensional interface such as a mouse or a touch screen.

To ease the labeling across frames, we have added a tracking feature to automatically label subsequent frames in the point cloud set. The tracking algorithm fits labeled points in $frame_i$ to a portion of points in $frame_{i+1}$, so that we can predict labels from one frame to the next. Figure 3 shows how three individual parts of the point cloud has been labeled, and then tracked from one frame to the next. We implemented two algorithms to fit labeled points. The first one is an Iterative Closest Point (ICP) algorithm [Besl et al. 1992], which iteratively detects the closest point in $frame_{i+1}$ for each labeled point in $frame_i$ and estimates the current best rotation

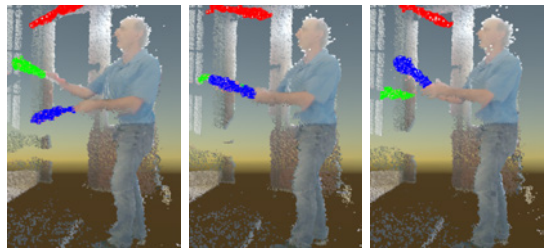


Figure 3: Three frames from the point cloud data. The juggling pins are labeled individually (red, green and blue) by the user in the first frame. The tracking algorithm is used to track points from one frame and automatically label them in the next.

and translation to update the positions of the labeled points. This iterative process terminates until a predetermined stop criteria satisfies, for example, the spatial difference between labeled points and matched points converges. The second algorithm follows the same pipeline as the first one, but is augmented with color support to improve tracking accuracy. Instead of detecting the spatially closest point, the second algorithm selects, for each query point, the point with the most similar color from k-nearest neighboring points as the match. We build a k-d tree [Brown 2014], which can be pre-calculated for each point cloud, to expedite the nearest neighbor search.

Our method is not limited to Kinect data, thus data of varying scale and resolution can be visualized and labeled. We also emphasize the potential of this application for modifying point clouds, such as deleting noisy points, or cutting out relevant parts of the data using the same type of labeling and selection method.

4 CONCLUSION

We have realized an application for visualizing and labeling point clouds in Virtual Reality. The application demonstrates that it is possible to label and inspect large sets of room-scale point clouds in a fast and intuitive way. With the application, it is possible to select and track parts of the data to speed up labeling in sequences of point clouds. We believe that is a step in the direction towards incorporating point clouds into an efficient production pipeline for realistic animated virtual reality content, and more high precision labeled point cloud data sets for a variety of other applications as well.

5 ACKNOWLEDGEMENTS

Thanks to the MIT Juggling Club for allowing us to record the point cloud data.

REFERENCES

Paul J Besl, Neil D McKay, et al. 1992. A method for registration of 3-D shapes. *IEEE Transactions on pattern analysis and machine intelligence* 14, 2 (1992), 239–256.

Russell A. Brown. 2014. Building a Balanced k-d Tree in O(kn log n) Time. *CoRR* abs/1410.5420 (2014). http://arxiv.org/abs/1410.5420

Semantic3d. 2016. Large-Scale Point Cloud Classification Benchmark. (2016). http://www.semantic3d.net/

Contribution E

Virtual reality inspection and painting with measured BRDFs

Virtual Reality inspection and painting with measured BRDFs

Alessandro Dal Corso
Technical University of Denmark

Jonathan Dyssel Stets
Technical University of Denmark

Andrea Luongo
Technical University of Denmark

Jannik Boll Nielsen
Technical University of Denmark

Jeppe Revall Frisvad
Technical University of Denmark

Henrik Aanæs
Technical University of Denmark

Figure 1: Pictures illustrating our VR demo application, with an in-game screenshot (left) and a picture of the setup (right). Paintbucket, table and lightbuld models ©TurboSquid.com, environment maps ©HDRMaps.com and ©Joost Vanhoutte.

CCS CONCEPTS

• **Computing methodologies** → **Virtual reality**;

KEYWORDS

Virtual Reality, material appearance

ACM Reference Format:
Alessandro Dal Corso, Jonathan Dyssel Stets, Andrea Luongo, Jannik Boll Nielsen, Jeppe Revall Frisvad, and Henrik Aanæs. 2017. Virtual Reality inspection and painting with measured BRDFs. In *Proceedings of SA '17 VR Showcase, Bangkok, Thailand, November 27-30, 2017,* 2 pages.
https://doi.org/10.1145/3139468.3139472

1 INTRODUCTION

This is a virtual reality (VR) painting application that enables the user to paint on 3D models with real measured materials, much like in the physical world. A scanned physical object can be imported into the VR application, and the user can paint on the surface of the object with a virtual hand-controlled paint brush. The user is presented with several paint buckets, each containing a material known from the physical world. These materials are measured bidirectional reflectance distribution functions (BRDFs) of real physical materials. The materials and objects present in VR are thus represented as they would be in the physical world, and the user can control both the environment lighting and a single light source. The application enables analog artists to apply their skills directly on a digital model and it enables engineers to directly inspect BRDFs in a fast and intuitive way. Figure 1 is a screenshot from the application showing models that have been painted with BRDFs.

2 MOTIVATION

The realization of our VR demo was mainly driven by two objectives. The first was to have an intuitive and practical way of inspecting and visualizing measured BRDFs. Our application enables the user to apply BRDFs to 3D objects in a simple way that mimics the action of painting in real life. The user can also interact with the environment and the objects through an intuitive interface. Some of the actions the user can perform consist of modifying the light source intensity and position, the environment map, and the position and orientation of the objects. All these features allow us to quickly visualize and inspect a chosen BRDF under different lighting conditions without having to interact with a complicated interface.

Our second objective was to have an application useful in industrial and artistic design [Wald et al. 2006]. We provide a new way for artists, both digital and analog, to transfer their skills to the 3D domain by painting materials directly on 3D models that they might have created or scanned from real objects. By providing a real-time rendering environment, we enable artists to immediately see the final results of their creations under different viewing perspectives and lighting conditions. With this VR application, we aim at moving

SA '17 VR Showcase, November 27-30, 2017, Bangkok, Thailand

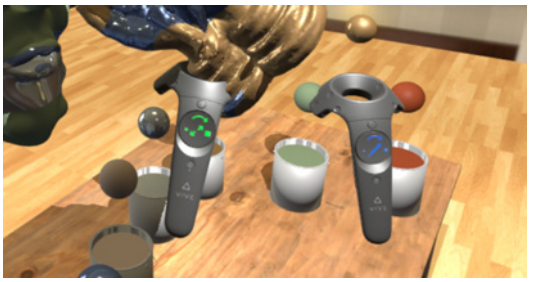


Figure 2: Interaction (left) and paintbrush (right) controller. Note the icons allowing to change the model and the brush size, respectively. See Figure 1 for copyright notice.

a first step towards the creation of a tool that will allow artists to apply materials and colors in a way that is similar to what they would do in the real world. Thus, we support the more traditional artistic workflow rather than inventing new artistic workflows for editing BRDFs [Colbert et al. 2006].

A user entering the virtual world will see a simple scene: a table, calibrated to match the position and height of a table present in the real world, a set of paint buckets containing different materials based on measured BRDFs, several 3D models, and a light-bulb. The user interacts with the scene through two handheld trackable controllers, one used as a painting tool and the other one as a grabbing and interaction tool. He/she can grab, move, rotate, and scale the 3D models with one hand and paint and select a different material with the other. The user can also move the position of the light-bulb and its intensity to get an immediate feedback on the appearance of their work. Furthermore, we use haptic feedback to enhance the interaction between user and objects in the scene. Our application thus creates a bridge between the digital and the physical domain, with an interface known from the physical world that the user is already familiar with. All these features help users immerse themselves and unfold their creativity in an environment similar to what they would see in a real artist's studio, and we have aimed to make this environment as interactive as possible.

3 DETAILS

Our demo is an HTC Vive setup (https://www.vive.com/), using the two provided controllers as interaction devices. One of the controllers is used for interaction, while the other acts as a paintbrush. The interaction is activated using the trigger button on the controller, and is used for grabbing and moving objects, including the light bulb above the table and the paint buckets. We use the small spheres on the left side of the table to change environment map. Finally, we have an undo button on the right to undo/redo the last action performed. Dipping the paintbrush into a bucket changes the BRDF that it will paint with. Based on the controller touchpads (see Figure 2), we also added interactions for changing object size (while grabbed), paintbrush size (on the paintbrush controller), and light intensity (while the light is grabbed).

We use measured BRDFs both from the MERL database [Matusik et al. 2003] and from our own laboratory. The scene has two light sources: an environment map and a movable point light in the form of a light bulb. The environment map contributes with background, reflections, and an ambient term. The ambient term is computed through standard spherical harmonics multiplied by the bihemispherical reflectance ρ of each measured BRDF, calculated in a preprocessing step using Monte Carlo integration. We multiply the environment map reflected color by ρ and by the factor $\min(1, \frac{f_{\max}}{C\rho_{\text{avg}}} - 1)$, where $f_{\max}$ is the peak value in the measured material, ρ_{avg} is the average of the three channels of ρ, and C is a user-defined constant. In case of a material with a strong reflection peak (such as a metallic paint), $\rho_{\text{avg}} \ll f_{\max}$ and the factor will be equal to one. In a case of a more diffuse material, $\rho_{\text{avg}} \approx f_{\max}$ and the reflection term will not be included. To paint the material, we intersect a sphere with the vertices of the model. For materials without a UV map, we write a material label on a per-vertex data structure. If a UV map is present, we first generate a secondary texture that maps vertex coordinates to UVs and then write the labels into a texture using the generated mapping.

4 USER FEEDBACK AND CONCLUSION

We invited an analog artist, a design engineer, and two 3D artists to test our application. They all found the interface intuitive to work with. Most noticeably, the analog artist used the full system without any previous VR experience after a one minute verbal instruction. This supports the objective of our application to enable transfer of artistic skills from the analog to the digital domain. Users noted the bulkiness of the controller compared to real-life painting tools such as a brush or a pencil. We accept this limitation of the system, hoping for smaller, lighter, or more customizable solutions in the future, like the stylus presented by Jackson and Keefe [2016].

Our VR painting application enables the user to paint on 3D models with measured BRDFs. Our users praised how our application is intuitive to use, and how it creates a bridge between the analog and digital skills. Furthermore, it enables a fast and intuitive way to inspect BRDFs under various lighting conditions.

Acknowledgements

Paint bucket, table, and light bulb models are courtesy of TurboSquid.com. Environment maps: studio, garden, and sunset forest from hdrmaps.com, lobby and night city square from Joost Vanhoutte. We use chrome, blue paint, gold paint and red fabric BRDFs from the MERL database. Ogre model courtesy of Keenan Crane. VIVE and related assets are property of HTC, Inc. and Valve Corporation. We would like to thank users Felicia Frisvad, Jon Murray Vinther, Sam Surplice, Christian Ahm for their valuable feedback.

REFERENCES

M. Colbert, S. Pattanaik, and J. Krivanek. 2006. BRDF-Shop: creating physically correct bidirectional reflectance distribution functions. *IEEE Computer Graphics and Applications* 26, 1 (Jan 2006), 30–36. https://doi.org/10.1109/MCG.2006.13

B. Jackson and D. F. Keefe. 2016. Lift-off: Using reference imagery and freehand sketching to create 3D models in VR. *IEEE Transactions on Visualization and Computer Graphics* 22, 4 (2016), 1442–1451. https://doi.org/10.1109/TVCG.2016.2518099

W. Matusik, H. Pfister, M. Brand, and L. McMillan. 2003. A Data-Driven Reflectance Model. *ACM Transactions on Graphics* 22, 3 (July 2003), 759–769.

I. Wald, A. Dietrich, C. Benthin, A. Efremov, T. Dahmen, J. Gunther, V. Havran, H. p. Seidel, and P. Slusallek. 2006. Applying Ray Tracing for Virtual Reality and Industrial Design. In *2006 IEEE Symposium on Interactive Ray Tracing*. 177–185. https://doi.org/10.1109/RT.2006.280229

Contribution F

Our 3D Vision Data-Sets in the Making

Our 3D Vision Data-Sets in the Making

H. Aanæs[1] K. Conradsen[1] A. Dal Corso[1] A. B. Dahl[1] A. Del Bue[2] M. Doest[1]
J. R. Frisvad[1] S. H. N. Jensen[1] J. B. Nielsen[1] J. D. Stets[1]
G. Vogiatzis[3]
[1] Technical University of Denmark [2] Istituto Italiano di Tecnologia [3] Aston University, UK

1. Introduction

Over the previous years, we have at the Section for Image Analysis and Computer Graphics at the Technical University of Denmark been working on generating high quality data sets for computer vision via our lab setup using a 6-axis industrial robot. This has provided a new data set aimed at feature matching [1, 4], and two data sets aimed at multiple view stereo [14, 16]. The resulting data sets are publicly available via `http://roboimagedata.compute.dtu.dk/`.

The evaluation of computer vision algorithms on these data sets has provided useful insights on realistic scenarios by setting a rigorous framework for evaluation. The results of these efforts have been well received by the community and the hardware and software platform associated with the robot is now well developed. We are currently in the process of making three new data sets aimed at 3D vision, with a special focus on the more challenging aspects, such as radiometry and the modelling of non-rigid objects. The construction of these data sets all leverage on our robotic setup's ability to produce ground truth camera and surface geometry, as briefly outlined in Section 2, and there is a great deal of commonality in the making of the data sets.

This abstract describes our current ongoing work on this data set construction for 3D vision. The data sets include:

1. A direct extension of our large multiple view stereo (MVS) data set [14], where we are now including transparent and semi transparent objects into the scenes, Section 3. A challenge in doing this is getting the ground truth geometry of the transparent objects.

2. A data set addressing the radiometric challenges in 3D vision as presented in Section 4 where we aim at extending our MVS data set by explicitly measure the bidirectional reflectance distribution function (BRDF) of the surfaces. This will have the additional feature to finally give a data set for evaluating photometric stereo with a ground truth.

3. An extension of our data set on feature matching to evaluate these algorithm with non-rigid objects, (Section 5) where we use actuators to make stop motion 3D data sets. This data set will also evaluate Non-rigid Structure from Motion (NRSfM) with realistic objects.

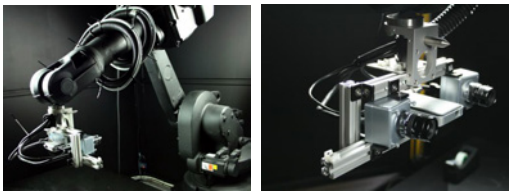


Figure 1. Photos of the 6-axis industrial robot mounted with two cameras and a projector. Cameras allow for MVS, and in conjunction with the projector SL provides ground truth point clouds.

2. Brief System Overview

Our experimental setup, cf. [1], is built around a 6-axis ABB IRB 1600 industrial robot, providing a flexible, precise, and highly repeatable camera pose. The robot is mounted with two Point Grey Grasshopper3 3376×2704 8-bit RGB cameras and a projector (for previously published datasets the cameras were 1600×1200 8-bit Point Grey Scorpion cameras). From each position ground truth surface point clouds are obtained using structured light (SL), and stereo images with a 32 cm baseline are captured with the camera pair. Five individually controlled 6500K LED tube lights allow for soft natural illumination of scenes from varying directions. Figure 1 shows the robot.

Previous evaluations of our system [14] have shown that the ground truth samples obtained through SL have good accuracy with a surface standard deviation of 0.14 mm. We expect similar or better performance in this data set. Positioning repeatability of the robot is very high, with a standard deviation of 0.0031 mm over two months.

Additional instruments used for generating the data include a CT (Computed Tomography) scanner for ground truth geometry of transparent objects (described in Section 3) and an illumination arch for controlled directional light-

Figure 2. Preliminary images from our data set. In the first row, three glass objects (sphere, bowl, teapot) with markers placed on. On the second row, three calibration and rendering tools part of the pipeline: a black and white checkerboard (coordinate estimation), an X-Rite ColorChecker® (color balance compensation) and a chrome sphere (environment light evaluation).

ing (described in Section 4).

3. Transparent Objects

Our goal is to extend our original MVS dataset to account for transparent objects where the focus is on reconstruction of geometry and appearance. Usually, the radiometric behavior of the objects used in 3D reconstructions is assumed diffuse and opaque. This leads to a number of simplifications that we cannot apply to transparent objects. In the case of transparent objects, refraction and reflection cause distortion effects that complicate reconstruction.

Previous methods acquire data sets useful for image-based rendering of a transparent object [18, 11]. However, these methods do not produce an actual triangle mesh and require special rendering techniques for reconstruction of the appearance of the transparent object. A survey on methods that do provide a triangle mesh is available [13]. In this survey, they note that CT scanning of refractive objects like glass is costly but straight forward. Thus, we use CT scanning to obtain ground truth geometry. Another way is to acquire shape and pose of a transparent object from motion [3]. In any case, there seems to be no data set, like the one we propose, which is useful for multiple view reconstruction of transparent objects.

3.1. Data

Our data set contains a set of multiple view HDR images of three glass objects with different radiometric properties (top row of Figure 2). We use a solid sphere, a bowl with lid (composed of two parts) and a teapot with multiple thin glass layers (composed of three parts). The walls of the bowl and the teapot have different thickness. A diffuse backdrop is provided for the objects. We have made this as a gradient checkerboard, so that one half of the squares varies in color from left to right, and the other half varies in color from top to bottom. In this way, we can see how light reflects, refracts and scatters through the objects. The refractive index of the glass objects will be estimated directly from the scanned images, or, if this is unsuccessful, by the use of a refractometer. We marked the objects with small black plastic spheres, in order to easily determine their position relative to the scene. In our data set, we also provide high-resolution triangle meshes generated from CT scans. We use these scans as ground truth for either geometrical reconstruction algorithms or physically based rendering algorithms for appearance modelling.

Our current data set creation procedure is as follows. First, we choose a sequence of camera positions and orientations for our industrial robot. The robot enables us to reproduce a given set of positions and orientations with a very high precision. Then, we capture a first set of images placing a black and white checkerboard in the scene. This is done to obtain the camera positions relative to the scanned objects and calculate camera parameters for the setup. Secondly, we scan a commercial color checker, which allows us to compensate for color channel alterations in the final images. Finally, we scan a chrome sphere to get an HDR environment map of the surroundings. We use the resulting map as a light source in our rendering algorithms [5], so we can simulate the resulting scene with high precision. After these three calibration steps, we can finally scan the glass objects using the same pre-defined path used for the calibration images.

Once compiled, we are planning to use this data set to verify that the radiometric models [9] properly describe the radiometric properties of the scene. To do this we plan to feed the ground truth of our data into a custom-built renderer based on the NVIDIA OptiX library [20], and see how well it reproduces the images. If successful, we have a validated computational model, which in principle we 'just' have to invert to do 3D reconstruction of transparent objects. Following this we plan at applying state of the art 3D reconstruction algorithms and quantify how far the state of the art has come toward solving this central 3D vision reconstruction problem.

4. BRDF measurements and Photometric Stereo

The radiometric behaviour of an object plays a crucial role in MVS. Often this behaviour has been ignored or at most assumed Lambertian. This allows for acceptable reconstructions of geometry, but often poor recovery of the reflectance. For more accurate MVS and reflectance capture, the BRDF of an object should be taken into account and this is a problem that receives a growing amount of

attention [24, 15]. Within the field of photometric stereo, the reflectance of an object is the key element in recovering surface normals and thereby indirectly the object's geometry. Also here, assumptions about reflectance are made, these include e.g. Lambertian behaviour [27] or isotropic BRDFs [12].

For both of the above areas, a multi-view data set having ground-truth reflectance behaviour would be of great value, and does, to our knowledge, not currently exist. We are therefore now working on a MVS data set where not only the ground-truth geometry is given, but also a densely sampled BRDF ground-truth for all materials in the scene. In the following, we will elaborate on the details of how this data set will be acquired and what it will include.

4.1. Concept

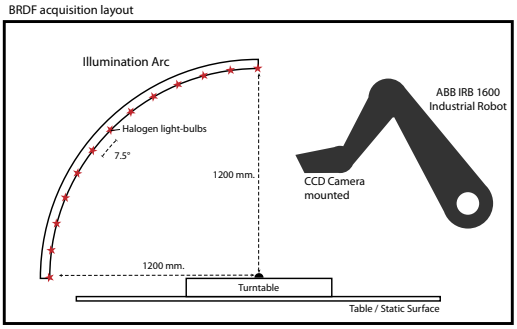


Figure 3. Schematic of BRDF capturing setup. Setup includes a 6-axis industrial robot holding a CCD (stereo) camera for view, and an arc in conjunction with a turntable for illumination.

Capturing the reflectance of a material generally requires four degrees of freedom: polar and azimuthal angle of illumination-direction, and polar and azimuthal angle of view-direction, $\rho(\omega_i, \phi_i, \omega_v, \phi_v)$. Utilizing our lab-facility's 6-axis industrial robot, mounted with a stereo-camera setup, all view directions (ω_v, ϕ_v) can effectively be captured. For illumination directions (ω_i, ϕ_i), we utilize an illumination arc and a rotation-table. The arc holds a range of halogen light-bulbs and is capable of covering the polar angle ϕ_i in 7.5° intervals. The rotation-table turns the target sample with a resolution of $< 1^\circ$, thus densely covering θ_i. Figure 3 shows a schematic of the BRDF capturing setup, and Figure 4 is a photo of an actual acquisition scene.

Using the above described setup, we intend to densely sample the BRDFs of a collection objects whose surfaces consist of one or a few, isotropic, BRDFs. The BRDFs of each material will be stored in the 3-dimensional Rusinkiewicz frame for isotropic BRDFs [21], as also done in the MERL database[17], although with a coarser resolution of 7.5° in each dimension. In conjunction with the densely sampled BRDFs, stereo images of scenes containing the sampled objects will be acquired for a wide range of directions. Objects will be of relatively low geometric complexity, and scenes will consist of one or more of the objects.

Figure 4. Capturing the BRDF of an object with known geometry. All illumination directions and view-directions are covered for each type material present on the object.

5. Non-Rigid Structure from Motion

Evaluating Non-rigid feature matching and NRSfM algorithms[1] in a quantitative manner has in the literature proven to be problematic. Deformations are inherently a dynamic process and subject to the physical properties of the objects in consideration. Thus, evaluating deformation modelling algorithms require a reasonable number of different objects and set of motions. Also, given the dynamic deformation objects might change their topology (e.g. stretching and tearing) and easily self-occluded some parts of the shape. For this reason, many approaches have provided several models that fit specific types of deformation, but that cannot comprise all of them. For this reason understanding the real performance of methods on realistic deformations is necessary to push forward advancements in this field.

The central problem of producing reference ground truth has been approached from many different angles. Several works compare their methods using synthetically generated images, as the true 3D geometry is readily available[26, 22, 19, 10]. Another popular approach is using MOCAP data, mainly human motion, for generating both test video sequence with 3D reference points [7, 26, 2, 10, 25]. Both falls short, as the former often lacks the complexity found in real-life scenes and the latter provides only a sparse set of reference points that are likely not to be possible to detect from images because of occlusions. As stated in [22, 8], there is a lack of and a need for a real-life NRSfM sequence with a dense 3D reference.

[1] A review on NRSfM methods, updated to 2010, can be found here: [23]

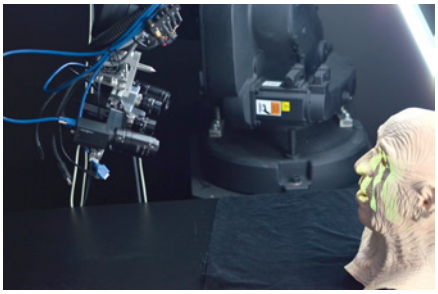


Figure 5. Robot arm carrying cameras for capturing stop motion frame and structured light data. A Gray code pattern is currently being projected onto the object.

We seek to remedy this situation by providing a video recording of real objects with dense 3D ground truth for each frame. It will be accomplished using a stop motion like animation techniques and structured light 3D scanning, combined in our unique recording setup.

5.1. Concept

We wish to simulate motion in a manner similar to stop motion animated films. Here a rigid object is moved into a certain pose, an image is taken, the object is slightly changed with a deformation, another image is taken etc. The result is a sequence that, when played at an interactive frame rate, provides the illusion of motion. We will apply the same principle here, in generating a benchmarking data set for NRSfM with ground truth.

Now one may ask, why not just record the motion using ordinary video format? After all, stop motion techniques does not properly reproduce motion blur artifacts that are present in standard recorded video sequences. Our approach has several significant advantages that greatly outweigh the loss of motion blur. Most importantly, we can obtain a 3D ground truth for each frame. After adjusting the object into its current frame position and acquiring an image for the stop motion sequence, we will perform a 3D scan using structured light. Utilizing gray code patterns we obtain a dense ground truth so obtaining both the image frame and a 3D reference for benchmarking and validation.

Another advantage is that we can obtain data from multiple views by acquiring images at different angles thus providing data for evaluating multi-view NRSfM (e.g. [6]). Furthermore, this procedure provides a great degree of control over both camera movement and object pose. As each frame is recorded independently, time in between becomes a non-issue.

Figure 6. Actuators for manipulating the geometry of the mask. The image of the mask has been superimposed on an image of the actuators, illustrating their functionality.

5.2. Implementation

Such data could be acquired by pure manual effort, however that would be extremely time consuming and error-prone. As such, a robotics solution is currently being developed with a the data acquisition procedure that is predictably and reproducibly implemented. In detail, a robotic arm move the camera and the projector needed for data acquisition and structured light scan. From this the view position can be determined with high precision and reproducibility. Figure 5 illustrates this setup.

Additionally, object deformation will also be automated and Figure 6 shows an example with an object where a mask resembling a human face is put on top of two actuators. Manipulating the actuators deforms the mask geometry, simulating facial movement. Similar results can be obtained with cloth, paper and other deformable materials.

6. Concluding Remarks

We have here presented our ongoing work on making high quality data sets for evaluating and developing methods for 3D vision. A motivation for doing this is that we see a need for this, especially with respect to making data sets that are large enough, so that it is possible to reasonably determine if differences in performance are a statistical fluke, or are in fact statistically significant.

By presenting our ongoing work in this forum, we hope to get valuable and constructive feedback on how these data sets in the making could be adapted to serve the needs of the computer vision communities as best possible.

References

[1] H. Aanæs, A. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *International Journal of Computer Vision*, 97(1):18–35, 2012.

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(7):1442–1456, 2011.

[3] M. Ben-Ezra and S. Nayar. What does motion reveal about transparency? In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 1025–1032, 2003.

[4] A. Dahl, H. Aanæs, and K. Pedersen. Finding the best feature detector-descriptor combination. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 318–325, 2011.

[5] P. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of ACM SIGGRAPH 98*, pages 189–198, 1998.

[6] A. Del Bue and L. Agapito. Stereo non-rigid factorization. *International Journal of Computer Vision*, 66(2):193–207, February 2006.

[7] J. Fayad, L. Agapito, and A. Del Bue. Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, pages 297–310. Springer, 2010.

[8] K. Fragkiadaki, M. Salas, P. Arbelaez, and J. Malik. Grouping-based low-rank trajectory completion and 3D reconstruction. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 55–63. Curran Associates, Inc., 2014.

[9] A. S. Glassner. Surface physics for ray tracing. In A. S. Glassner, editor, *An Introduction to Ray Tracing*, chapter 4, pages 121–160. Academic Press Ltd., London, UK, 1989.

[10] P. F. U. Gotardo and A. M. Martinez. Kernel non-rigid structure from motion. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 802–809. IEEE, 2011.

[11] T. Hawkins, P. Einarsson, and P. E. Debevec. A dual light stage. *Rendering Techniques 2005 (Proceedings of EGSR 2005)*, pages 91–98, 2005.

[12] M. Holroyd, J. Lawrence, G. Humphreys, and T. Zickler. A photometric approach for estimating normals and tangents. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2008)*, 27(5):133, 2008.

[13] I. Ihrke, K. N. Kutulakos, H. Lensch, M. Magnor, and W. Heidrich. Transparent and specular object reconstruction. *Computer Graphics Forum*, 29(8):2400–2426, 2010.

[14] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 406–413, 2014.

[15] H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo beyond Lambert. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I:171–178. IEEE, 2003.

[16] S. Kim, H. Aanæs, A. Dahl, K. Conradsen, R. Jensen, and S. Kim. Multiple view stereo by reflectance modeling. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, 2012.

[17] W. Matusik, H. Pfister, M. Brand, and L. McMillan. A data-driven reflectance model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2003)*, 22(3):759–769, 2003.

[18] W. Matusik, H. Pfister, R. Ziegler, A. Ngan, and L. McMillan. Acquisition and rendering of transparent and refractive objects. pages 267–278, 2002.

[19] S. I. Olsen and A. Bartoli. Implicit non-rigid structure-from-motion with priors. *Journal of Mathematical Imaging and Vision*, 31(2-3):233–244, 2008.

[20] S. G. Parker, J. Bigler, A. Dietrich, H. Friedrich, J. Hoberock, D. Luebke, D. McAllister, M. McGuire, K. Morley, A. Robison, and M. Stich. OptiX: a general purpose ray tracing engine. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2010)*, 29(4):66:1–66:13, July 2010.

[21] S. Rusinkiewicz. A new change of variables for efficient BRDF representation. In *Rendering Techniques (Proceedings of EGWR 1998)*, June 1998.

[22] C. Russell, J. Fayad, and L. Agapito. Dense non-rigid structure from motion. In *Proceedings of 3DIMPVT 2012*, pages 509–516. IEEE, 2012.

[23] M. Salzmann and P. Fua. Deformable surface 3d reconstruction from monocular images. *Synthesis Lectures on Computer Vision*, 2(1):1–113, 2010.

[24] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 519–528. IEEE, 2006.

[25] L. Tao and B. J. Matuszewski. Non-rigid structure from motion with diffusion maps prior. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 1530–1537. IEEE, 2013.

[26] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 2761–2768, 2010.

[27] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):191139, 1980.

Contribution G

3D heatmap in marketing research and marketing practice - validation of an integrating model

Not included in the public version of this thesis.

Contribution H

Learning Refraction with Convolutional Neural Networks

Technical Note
Learning Refraction with Convolutional Neural Networks

Note: *The following is an ongoing research project, and the results presented in this technical note are preliminary. The work has been carried out during the PhD studies and is added because it has high relevance to the other contributions presented in this thesis. The note is written by Jonathan Dyssel Stets, and the project is carried out in collaboration with Manmohan Chandraker and Zhengqin Li, University of California, San Diego and Jeppe Revall Frisvad, Technical University of Denmark.*

Refractive objects, such as glass, crystal, ice, and some plastics, all have in common that light can pass through them. In this note, a refractive object is mainly going to be a perfect glass, i.e., without air bubbles, surface scratches, or similar.

Refraction occurs when light changes from one medium to another, and the angle depends on the propagation speed in the medium. Snell's Law describes this relationship. If the incidence angle is very large, then the ray will reflect instead of refracting. This is known as a total internal reflection, and the angle at which this occurs is denoted the critical angle. But it can also be the case that both reflection and refraction occur simultaneously, which is defined by the Fresnel equation.

Because of these different types of interactions, glass has a unique appearance, which almost exclusively is determined by its surroundings. Consequently, it proves to be a complex task for visual systems to cope with glass objects [1], while humans seem to be able to understand the geometry of glass by visual inspection. Convolutional Neural Networks have proven to perform quite well on a broad spectrum of computer vision tasks, such as segmentation or depth estimation. Such an approach also proves to require a large amount of data, but rendering of synthetic training images can be a way to solve this need for data [4]. Consequently, in this technical note, we use a learning-based approach to investigate the possibility of enabling a computer to understand the geometry of glass, using synthesized data.

1 Data

When working with deep neural networks, there is a need for extensive training data. This can be a complicated task as some types of data can be hard to acquire and when it comes to methods for acquiring geometry of glass, the options are somewhat limited. We need both photos of the glass and geometric ground truth information. A common method is to coat the glass with a diffuse spray, like chalk, and then scan it using a visible light-based approach. This is, however, a slow procedure, and the glass objects need cleaning and in worst case risk of losing its original appearance by scratches. A non-intrusive method is to use a Computed Tomography (CT) scanner, but this scanning and reconstruction procedure is cumbersome.

Instead, we choose to create synthetic data using the OptiX Raytracing Engine [7] from NVIDIA. A customizable rendering tool gives complete control over what data to generate, and images sets can be generated relatively fast compared to a manual acquisition.

We use a set of high dynamic range environment maps and a set of 3D models. The models are a combination of models from Shapenet [2], and randomly generated shapes. A demonstration of the types of data generated by the render is shown in Figure 1. When generating data, we vary parameters such as rotation of environment map, rotation, and translation of objects and refractive index.

2 Training

The data is intended for a range of experiments, such as depth estimation, normal estimation, optical flow, and segmentation. All related to estimating the geometry of the objects.

As a first attempt, we wish to estimate depth from a single image. This has been done for real-life scenes [5, 6], which proves to work on a depth dataset recorded with a Microsoft Kinect. For our project, we use an encoder-decoder network architecture based on VGG16 [8] pretrained on ImageNet [3] for both segmentation and depth estimation. The segmentation network is trained with the mask image, and the depth network is trained with relative depth maps generated by the renderer. First, the network is trained on images with non-refractive objects - i.e., a textured version of Figure 1(f). Then it is refined with images of refractive objects as shown in Figure 1(a).

3 Preliminary Results

The network prediction is here demonstrated on two images shown in Figure 2. The preliminary results for the mask predictions and depth are shown in Figure 3 and Figure 4 respectively.

The results indicate that depth can be learned for synthetically generated refractive objects. Potentially, other geometric clues such as optical flow and thickness

(a) Glass image. (b) Reflected only. (c) Refracted only.

(d) MPP Coded. (e) Normal map. (f) Lambertian.

(g) Mask. (h) Depth map. (i) Thicnkess map.

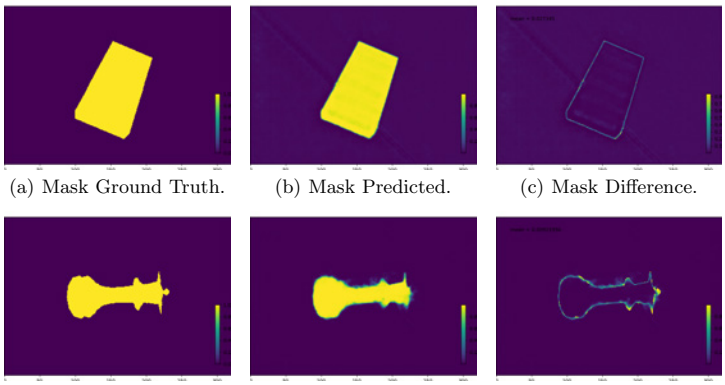
(j) Environment map.

Fig. 1. The different types of rendered data generated for the dataset. 1(d) utilizes a coded environment map, so this image shows a map of the Most Probable Path of rays in 1(a), where colors correspond to XYZ locations on the environment map.

can also be learned, enabling a technique to estimate geometry of refractive objects. Ideally, the renderings are photorealism enough so that the network also can predict from images of real glass objects. In that case, it would be ideal to render more everyday-like scenes with glass objects to be used as training data.

(a) Diffuse Test Image (b) Refractive Test Image

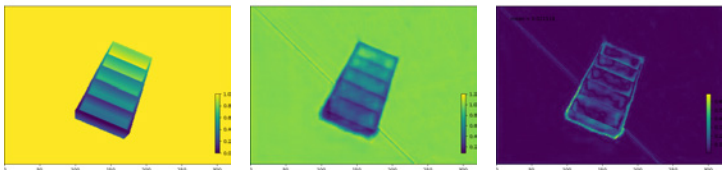
Fig. 2. The test images used to demonstrate the diffuse and refractive training results respectively.



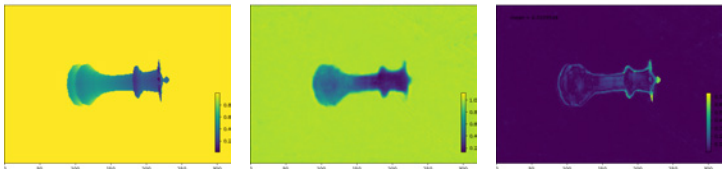
(a) Mask Ground Truth. (b) Mask Predicted. (c) Mask Difference.

(d) Mask Ground Truth. (e) Mask Predicted. (f) Mask Difference.

Fig. 3. Mask images: Top row is the diffuse object from Figure 2(a), and bottom row is the refractive object from Figure 2(b).



(a) Depth Ground Truth. (b) Depth Predicted. (c) Depth Difference.



(d) Depth Ground Truth. (e) Depth Predicted. (f) Depth Difference.

Fig. 4. Depth images: Top row is the diffuse object from Figure 2(a), and bottom row is the refractive object from Figure 2(b).

Bibliography

[1] Moshe Ben-ezra and Shree K. Nayar. S.: What does motion reveal about transparency. 2013. doi: 10.1.1.360.5865.

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

[5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[7] Steven G. Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, and Martin Stich. Optix: A general purpose ray tracing engine. *ACM Trans. Graph.*, 29(4):66:1–66:13, July 2010. ISSN 0730-0301. doi: 10.1145/1778765.1778803.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Bibliography

[ACD+15] Henrik Aanæs, Knut Conradsen, Alessandro Dal Corso, Anders Bjorholm Dahl, A. Del Bue, Mads Emil Brix Doest, Jeppe Revall Frisvad, Sebastian Hoppe Nesgaard Jensen, Jannik Boll Nielsen, Jonathan Dyssel Stets, and George Vogiatzis. Our 3d vision data-sets in the making, 2015.

[AD15] Henrik Aanæs and Anders Bjorholm Dahl. Accuracy in robot generated image data sets. *Lecture Notes in Computer Science*, pages 472–479, 2015.

[AJV+16] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016.

[App68] Arthur Appel. Some techniques for shading machine renderings of solids. In *Proceedings of the April 30–May 2, 1968, Spring Joint Computer Conference*, AFIPS '68 (Spring), pages 37–45, New York, NY, USA, 1968. ACM.

[Ash02] Michael Ashikhmin. A tone mapping algorithm for high contrast images, 2002.

[BeKN13] Moshe Ben-ezra and Shree K. Nayar. S.: What does motion reveal about transparency. 2013.

[BETVG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[BK04] Julia F Barrett and Nicholas Keat. Artifacts in ct: recognition and avoidance. *Radiographics*, 24(6):1679–1691, 2004.

[Boj09] Agnieszka Aga Bojko. Informative or misleading? heatmaps deconstructed. In *International Conference on Human-Computer Interaction*, pages 30–39. Springer, 2009.

[Bou] Jean-Yves Bouguet. Camera calibration toolbox for matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/`. [Online; accessed 2018-01-29].

[Bra00] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[Deb08] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *ACM SIGGRAPH 2008 Classes*, SIGGRAPH '08, pages 32:1–32:10, New York, NY, USA, 2008. ACM.

[Deb12] Paul Debevec. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia*, 2(4), 2012.

[DM97] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, pages 369–378, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[DSL+17] Alessandro Dal Corso, Jonathan Dyssel Stets, Andrea Luongo, Jannik Boll Nielsen, Jeppe Revall Frisvad, and Henrik Aanæs. *Virtual reality inspection and painting with measured BRDFs*. 2017.

[Duc07] Andrew Duchowski. *Eye Tracking Techniques*, pages 51–59. Springer London, London, 2007.

[EMU17] Gabriel Eilertsen, RK Mantiuk, and Jonas Unger. A comparative review of tone-mapping algorithms for high dynamic range video. In *Computer Graphics Forum*, volume 36, pages 565–592. Wiley Online Library, 2017.

[EPF14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[EPF+17] Eyþór Rúnar Eiríksson, David Bue Pedersen, Jeppe Revall Frisvad, Linda Skovmand, Valentin Heun, Pattie Maes, and Henrik Aanæs. Augmented reality interfaces for additive manufacturing. *Lecture Notes in Computer Science*, 10269:515–525, 2017.

[EWPA16] E.R. Eiríksson, J. Wilm, D. B. Pedersen, and H. Aanæs. Precision and accuracy parameters in structured light 3-d scanning. volume XL-5/W8, pages 7–15, Gottingen, 2016. Copernicus GmbH. Copyright - Copyright Copernicus GmbH 2016; Last updated - 2018-02-05.

[FB81] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[FLM92] O. D. Faugeras, Q. T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In G. Sandini, editor, *Computer Vision — ECCV'92*, pages 321–334, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg.

[FP10] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010.

[Gen11] Jason Geng. Structured-light 3d surface imaging: a tutorial. *Adv. Opt. Photon.*, 3(2):128–160, Jun 2011.

[Gla89] Andrew S Glassner. *An introduction to ray tracing*. Elsevier, 1989.

[GPBS14] Leonardo Gomes, Olga Regina Pereira Bellon, and Luciano Silva. 3d reconstruction methods for digital preservation of cultural heritage: A survey. *Pattern Recognition Letters*, 50:3–14, 2014.

[GRS10] I. Gibson, D. W. Rosen, and B. Stucker. Additive manufacturing technologies: Rapid prototyping to direct digital manufacturing. *Additive Manufacturing Technologies: Rapid Prototyping To Direct Digital Manufacturing*, pages 1–459, 2010.

[GTGB84] Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. Modeling the interaction of light between diffuse surfaces. *SIGGRAPH Comput. Graph.*, 18(3):213–222, January 1984.

[Har93] Richard I. Hartley. Euclidean reconstruction from uncalibrated views. In *Applications of Invariance in Computer Vision*, pages 237–256. Springer-Verlag, 1993.

[HJ10] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2010.

[HNA+11] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[HRV97] B. Hill, Th. Roger, and F. W. Vorhagen. Comparative analysis of the quantization of color spaces on the basis of the cielab color-difference formula. *ACM Trans. Graph.*, 16(2):109–154, April 1997.

[HS88] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[HS97] Janne Heikkila and Olli Silven. A four-step camera calibration procedure with implicit image correction. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 1106–1112. IEEE, 1997.

[HTC] HTC. Vive. `https://www.vive.com/`. [Online; accessed 2018-02-10].

[HZ03] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[iMo] iMotions. imotions biometric research platform. `https://imotions.com/`. [Online; accessed 2018-02-10].

[INKPAL+13] Ivo Ihrke, Kiriakos N. Kutulakos, Hendrik P. A. Lensch, Marcus Magnor, and Wolfgang Heidrich. Transparent and specular object reconstruction. 2013.

[Jak] Wenzel Jakob. Mitsuba. `https://www.mitsuba-renderer.org`. [Online; accessed 2018-01-31].

[JGBG17] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *CoRR*, abs/1704.05519, 2017.

[JSS+17] Rasmus Ramsbøl Jensen, Jonathan Dyssel Stets, Seidi Suurmets, Jesper Clement, and Henrik Aanæs. *Wearable Gaze Trackers: Mapping Visual Attention in 3D*, pages 66–76. Springer, 2017.

[KS11] Jonathan Kelly and Gaurav S Sukhatme. Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration. *The International Journal of Robotics Research*, 30(1):56–79, 2011.

[KVD91] Jan J Koenderink and Andrea J Van Doorn. Affine structure from motion. *JOSA A*, 8(2):377–385, 1991.

[KW13] Kuno Kurzhals and Daniel Weiskopf. Space-time visual analytics of eye-tracking data for dynamic stimuli. *Ieee Transactions on Visualization and Computer Graphics*, 19(12):2129–2138, 2013.

[LCR01] M Ronnier Luo, Guihua Cui, and B Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26(5):340–350, 2001.

[LHZB+16] Yaron Lipman, Richard Hao Zhang, Sofien Bouaziz, Andrea Tagliasacchi, and Mark Pauly. Sparse iterative closest point. 2016.

[Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[Mat] MathWorks. Computer vision system toolbox. `https://se.mathworks.com/products/computer-vision.html`. [Online; accessed 2018-01-29].

[MHB14] Michael Maurus, Jan Hendrik Hammer, and Jürgen Beyerer. Realistic heatmap visualization for interactive analysis of 3d gaze data. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14, pages 295–298, New York, NY, USA, 2014. ACM.

[Mic] Microsoft. Microsoft hololens. `https://www.microsoft.com/hololens`. [Online; accessed 2018-02-06].

[MPBM03] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A data-driven reflectance model. *ACM Transactions on Graphics*, 22(3):759–769, July 2003.

[MT12] D. Moreno and G. Taubin. Simple, accurate, and robust projector-camera calibration. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission*, pages 464–471, Oct 2012.

[Nic65] Fred E. Nicodemus. Directional reflectance and emissivity of an opaque surface. *Appl. Opt.*, 4(7):767–775, Jul 1965.

[Nie16] Jannik Boll Nielsen. On practical sampling of bidirectional reflectance, 2016.

[NOCM12] A. Y. C. Nee, S. K. Ong, G. Chryssolouris, and D. Mourtzis. Augmented reality applications in design and manufacturing. *Cirp Annals-manufacturing Technology*, 61(2):657–679, 2012.

[NSL+17] Jannik Boll Nielsen, Jonathan Dyssel Stets, Rasmus Ahrenkiel Lyngby, Henrik Aanæs, Anders Bjorholm Dahl, and Jeppe Revall Frisvad. *A variational study on BRDF reconstruction in a structured light scanner*, pages 143–152. IEEE, 2017.

[Ocu] Oculus. Oculus rift. `https://www.oculus.com/`. [Online; accessed 2018-02-10].

[PB06] Alex Poole and Linden J Ball. Eye tracking in hci and usability research. *Encyclopedia of human computer interaction*, 1:211–219, 2006.

[PBD+10] Steven G. Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, and Martin Stich. Optix: A general purpose ray tracing engine. *ACM Trans. Graph.*, 29(4):66:1–66:13, July 2010.

[Pfe12] Thies Pfeiffer. Measuring and visualizing attention in space with 3d attention volumes. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pages 29–36, New York, NY, USA, 2012. ACM.

[PJH16] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation.* Morgan Kaufmann, 2016.

[Pla] Playstation. Playstation vr. `https://www.playstation.com/playstation-vr/`. [Online; accessed 2018-02-10].

[pM07] Oleg Špakov and Darius Miniotas. Visualization of eye gaze data using heat maps. 115, 01 2007.

[Proa] Tobii Pro. Tobii pro glasses 2. `https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/`. [Online; accessed 2018-02-05].

[Prob] Blender Project. Cycles. `https://www.cycles-renderer.org/`. [Online; accessed 2018-01-31].

[PSF+13] Lucas Paletta, Katrin Santner, Gerald Fritz, Albert Hofmann, Gerald Lodron, Georg Thallinger, and Heinz Mayer. A computer vision system for attention mapping in slam based 3d models. 2013.

[RHD+10] Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.

[Rom95] Lois E Romans. *Introduction to Computed Tomography*. Williams & Wilkins, 1995.

[RSSF02] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, July 2002.

[SDA11] Jonathan Dyssel Stets, Anders Lindbjerg Dahl, and Henrik Aanæs. 3d surface scanner using structured light & industrial robot. 2011.

[SDN+17] Jonathan Dyssel Stets, Alessandro Dal Corso, Jannik Boll Nielsen, Rasmus Ahrenkiel Lyngby, Sebastian Hoppe Nesgaard Jensen, Jakob Wilm, Mads Brix Doest, Carsten Gundlach, Eythor Runar Eiriksson, Knut Conradsen, Anders Bjorholm Dahl, Jakob Andreas Bærentzen, Jeppe Revall Frisvad, and Henrik Aanæs. Scene reassembly after multimodal digitization and pipeline evaluation using photorealistic rendering. *Applied Optics*, 56(27):7679–7690, 2017.

[SFPL10] Joaquim Salvi, Sergio Fernandez, Tomislav Pribanic, and Xavier Llado. A state of the art in structured light patterns for surface profilometry. *Pattern recognition*, 43(8):2666–2680, 2010.

[SNDM11] Sophie Stellmach, Lennart Nacke, Raimund Dachselt, and Otto-von-guericke-universität Magdeburg. Advanced gaze visualizations for three-dimensional virtual environments. 2011.

[SS01] Linda G. Shapiro and George C Stockman. *Computer Vision*. Prentice Hall, 2001.

[SSGC17] Jonathan Dyssel Stets, Yongbin Sun, Scott W. Greenwald, and Wiley Corning. *Visualization and labeling of point clouds in virtual reality*. 2017.

[SWD05] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

[Sze11] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2011.

[TKM+] Blascheck T., Kurzhals K., Raschke M., Burch M., Weiskopf D., and Ertl T. Visualization of eye tracking data: A taxonomy and survey. *Computer Graphics Forum*, 36(8):260–284.

[TMHF99] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999.

[TR93] J. Tumblin and H. Rushmeier. Tone reproduction for realistic images. *IEEE Computer Graphics and Applications*, 13(6):42–48, Nov 1993.

[UWP06] Christiane Ulbricht, Alexander Wilkie, and Werner Purgathofer. Verification of physically based rendering algorithms. *Computer Graphics Forum*, 25(2):237–255, 2006.

[WBE+13] Sabine Webel, Uli Bockholt, Timo Engelke, Nirit Gavish, Manuel Olbrich, and Carsten Preusche. An augmented reality training platform for assembly and maintenance skills. *Robotics and Autonomous Systems*, 61(4):398–403, 2013.

[WBG+12] M.J. Westoby, J. Brasington, N.F. Glasser, M.J. Hambrey, and J.M. Reynolds. 'structure-from-motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300 – 314, 2012.

[Whi80] Turner Whitted. An improved illumination model for shaded display. *Commun. ACM*, 23(6):343–349, June 1980.

[Wil16] Jakob Wilm. *Real Time Structured Light and Applications*. PhD thesis, Technical University of Denmark (DTU), 2016.

[ZDES16] Longyu Zhang, Haiwei Dong, and Abdulmotaleb El Saddik. From 3d sensing to printing: A survey. *Acm Transactions on Multimedia Computing Communications and Applications*, 12(2):27, 2016.

[Zha00] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.